DOCUMENT RESUME

ED 068 132 LI 003 928

AUTHOR	Williams, Martha E.; And Others
TITLE	Educational and Commercial Utilization of a Chemical Information Center, Four Year Summary.
INSTITUTION	Illinois Inst. of Technology, Chicago. Computer Search Center.
SPONS AGENCY	National Science Foundation, Washington, D.C. Office of Science Information Services.
REPORT NO	C6156-18
PUB DATE	30 Jul 72
NOTE	425p.;(0 References)
EDRS PRICE	MF-$0.65 HC-$16.45
DESCRIPTORS	Annual Reports; *Chemistry; Computers; Data Bases; *Information Centers; *Information Retrieval; Information Services; Search Strategies; Use Studies

ABSTRACT

The major objective of the IITRI Computer Search Center is to educate and link industry, academia, and government institutions to chemical and other scientific information systems and sources. The Center was developed to meet this objective and is in full operation providing services to users from a variety of machine-readable data bases with minimal restrictions and a high degree of flexibility. A new modular machine-independent PL/1 software system was developed for handling virtually any bibliographic-type data base. Research is conducted and statistics maintained to continuously study, monitor, and improve Center components, including data bases, profiles, systems, personnel functions, and user services. Education and training is provided through seminars, workshops in profile preparation, and a graduate course in "Modern Techniques in Chemical Information." The educational and marketing efforts familiarize users and potential users with the many advantages of computerized retrieval, which are the raison d'etre for the Center, including: access to wide coverage; thoroughness of search; consistency of search; interdisciplinariness of data bases; high recall; speed of search; regularity of information dissemination; timeliness; automated personal file preparation and maintenance; and cost effectiveness. (A number of pages may be illegible.) (Author/SJ)

ED 068132

U.S. DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

FOUR-YEAR SUMMARY

EDUCATIONAL AND COMMERCIAL UTILIZATION OF A CHEMICAL INFORMATION CENTER

June 25, 1968 to June 25, 1972
Report No. C6156-18
Contract No. NSF-C554

To

National Science Foundation
Office of Science Information Service
Washington, D.C. 20550

Prepared by

IIT Research Institute
Computer Search Center
10 W. 35th Street
Chicago, Illinois 60616

Martha E. Williams
Peter B. Schipma
Scott E. Preece
David S. Becker
Patricia A. Llewellen
Alan K. Stewart

FILMED FROM BEST AVAILABLE COPY

July 30, 1972

LI 003 928

FOREWORD

This Summary Report No. IITRI-C6156-18 entitled "Four Year Summary--Educational and Commercial Utilization of a Chemical Information Center" summarizes work carried out on IITRI Project C6156 for the period June 25, 1968 to June 25, 1972. The project was funded by the National Science Foundation under Contract NSF-C554 and was monitored via the NSF Office of Science Information Service.

The project leader throughout the four years was Martha E. Williams, the principal investigator until April 1971 was Eugene Schwartz, and the programming coordinator throughout the time period was Peter B. Schipma.

Contributions to this report were made by Martha E. Williams, Peter B. Schipma, Scott E. Preece, David S. Becker, Patricia A. Llewellen, and Alan K. Stewart.

We would like to acknowledge the significant contributions to the project made by Eugene S. Schwartz in the area of system design, by Barbara M. Louthan in the area of programming the logic evaluation, and by Elaine Onderisin in originating the Least Common Bigram search technique. The project has been carried out as a team effort and significant inputs to the system design, design of programming module functions, operational procedures, and user requirements were made by all of the professional staff. We would like to acknowledge the former staff members Barbara Boone, John North, Henry Saxe, and Allan Shafton whose efforts have contributed to the success of the program. Finally, we would like to thank Arline Finnegan whose efforts as the CSC technician in handling output and maintaining records have provided essential support to the Center.

Prepared by
Martha E. Williams

Martha E. Williams
Manager
Information Sciences

ABSTRACT

FOUR-YEAR SUMMARY
EDUCATIONAL AND COMMERCIAL UTILIZATION OF A CHEMICAL INFORMATION CENTER

The major objective of the IITRI Computer Search Center (CSC) is to educate and link industry, academia, and government institutions to chemical and other scientific information systems and sources. The CSC was developed to meet this objective and is in full operation providing services to users from a variety of machine-readable data bases with minimal restrictions and a high degree of flexibility. A new modular machine-independent PL/1 software system was developed for handling virtually any bibliographic-type data base. CSC's transferable programs have run at fifteen different computer facilities with different: hardware, computer models, versions of OS, peripherals, and releases of the PL/1 compiler. All data bases are converted by a preprocessor to a standard IITRI format which employs a directory and character string type of file structure and are searched by a software system that employs the novel IITRI-developed Least Common Bigram search screen technique.

User oriented profile features include: full free form Boolean logic with any degree of nesting; search terms may be any data element on a data base; search terms may be single words, multi-word terms, phrases, or term fragments; full truncation capabilities; option for sorting output by author, citation number, or weight; and options for sorting output by author, on 5" x 8" cards, multilith masters, paper, magnetic tape, or COM. User aids were developed for each data base to assist in profile development and monitoring. They include: a Search Manual, data base oriented supplements to the Search Manual, Truncation Guides, term frequency lists, KLIC Indexes, and Search Term Frequency/Issue lists for each profile.

Research is conducted and statistics maintained to continuously study, monitor, and improve Center components including data bases, profiles, systems, personnel functions, and user services.

Education and training is provided through seminars, workshops in profile preparation, and a graduate course in "Modern Techniques in Chemical Information". The educational and marketing efforts familiarize users and potential users with the many advantages of computerized retrieval, which are the raison d'être for the center, including: access to wide coverage; thoroughness of search; consistency of search; interdisciplinariness of data bases; high recall; speed of search; regularity of information dissemination; timeliness; automated personal file preparation and maintenance; and cost effectiveness.

TABLE OF CONTENTS

Page

1. INTRODUCTION 1
   - 1.1 Report Description 1
   - 1.2 History and Background 3
2. COMPUTER SEARCH CENTER DESIGN AND DEVELOPMENT 6
   - 2.1 Objectives 6
   - 2.2 Design and Development 7
   - 2.3 Programming Language Selection 9
   - 2.4 Computer 11
   - 2.5 Modularity and Program Modules 12
   - 2.6 Transferability--Machine and Installation Independence 15
3. SERVICES 17
   - 3.1 SDI 17
   - 3.2 Retrospective Searches 18
   - 3.3 Private Libraries 28
   - 3.4 Software Installation 29
4. PROFILE PREPARATION AND MODIFICATION 31
   - 4.1 Profile Forms 31
   - 4.2 Profile Options and Features 35
     - 4.2.1 Terms 35
     - 4.2.2 Truncation 36
     - 4.2.3 Linking or Grouping of Terms 40
     - 4.2.4 Logic 40
     - 4.2.5 Weights 43
   - 4.3 Profile Format 43
   - 4.4 Profile Modification 45
5. SOFTWARE SYSTEM 48
   - 5.1 Data Base File Structure and Preprocessor System 56
   - 5.2 Format Conversion Programs 61

TABLE OF CONTENTS (cont.)

Page

5.2.1 Variability of Data Base Format 61
5.2.2 Data Base Documentation 62
5.2.3 Programming 63
5.2.4 Status 65
5.3 Program Descriptions 66
5.3.1 DBCOPY 66
5.3.2 FORCON 66
5.3.3 IFCOPY 67
5.3.4 DKEDIT 67
5.3.5 MINIPUP 68
5.3.6 INPUTR 68
5.3.7 SEARCH 69
5.3.8 DBCARD 70
5.3.9 DB-OCP 71
5.3.9.1 0CP-0CP1 71
5.3.9.2 0CP-S0RT 1 71
5.3.9.3 0CP-S0RT 2 72
5.3.9.4 0CP-0CP2 72
5.3.9.5 OCP-MICRO 73
5.3.10 HITTER 73
5.3.11 STIXA 73
5.3.12 PLSXT 74
5.4 Core Storage Requirements 75
5.5 System Files and File Structure 77
5.5.1 CA-C0ND.SDF1 77
5.5.2 DBOCPPRT 77
5.5.3 UFDBvvii 78
5.5.4 IFDBvvii 78
5.5.5 IFRFDBvv 79
5.5.6 M.CRO.OUTPUT 79
5.5.7 PCSCBKUP 79
5.5.8 PCSCCARD 80
5.5.9 PCSCCITS 81

TABLE OF CONTENTS (cont.)

Page

5.5.10 PCSCFMCT 81
5.5.11 PCSCHEAD 82
5.5.12 PCSCHITS 82
5.5.13 PCSCLOGC 83
5.5.14 PCSCOPLG 84
5.5.15 PCSCPASS 85
5.5.16 PCSCSRTD 85
5.5.17 PCSCSTHD 86
5.5.18 PCSCSTHT 86
5.5.19 PCSCSTFD 86
5.5.20 PCSCTERM 87
5.5.21 POUTPRNT_1 88
5.5.22 POUTPRNT_2 88
5.5.23 PRIMASTR, PRISRTOT, PRIWKOUT 88

5.6 Search Algorithm Comparison 90
5.6.1 Introduction 90
5.6.2 Term-to-Citation Algorithm 93
5.6.3 Basic Citation-to-Term Algorithm 93
5.6.4 Basic Citation-to-Term via Initial Bigram 95
5.6.5 Basic Citation-to-Term via Initial Trigram 96
5.6.6 Basic Citation-to-Term via Least Common Bigram 97
5.6.7 Summary 101

5.7 Logic Evaluation 102
5.7.1 Early Operator Reverse Polish 102
5.7.2 Tree Logic 104
5.7.3 Modified EORP 105

5.8 Private Libraries System 106
5.8.1 Private Libraries System--General 106
5.8.2 Software System 107

TABLE OF CONTENTS (cont.)

Page

5.8.2.1 PLSXT 108
5.8.2.2 PLSST 109
5.8.2.3 Special Conversion Programs 109
5.8.2.4 PRILIB 109
5.8.3 User Interaction with Files--Commands 110
5.8.4 Libraries 112
5.8.5 Use of the System 113
6. DATA BASES--CHARACTERISTICS, STATISTICS, AND COMPARISONS 115
6.1 Data Base Characteristics 115
6.1.1 Number of Citations per Issue 115
6.1.2 Statistics on Length of Citations, Data Fields per Citation and Key Words per Citation 116
6.1.3 Percent Occurrence of Data Types 116
6.1.4 Average Length of Data Entries by Data Types 124
6.2 Data Base Term and Character Occurrences 124
6.2.1 Term Frequency 124
6.2.2 Type:Token Ratios 134
6.2.3 Key-Letter-in-Context Listings 138
6.2.4 Bigram Frequencies 142
6.2.5 Single Character Frequencies 142
6.3 Data Base Terminology Variation 146
7. USER AIDS 150
7.1 Search Manual and Supplemental Guide 150
7.2 Key-Letter-in-Context (KLIC) Indexes 151
7.3 Term Frequency Lists 153
7.4 Truncation Guide and Standard Truncation 154
8. USER EVALUATION AND FEEDBACK 159
9. EDUCATION--USER LIAISON 167
9.1 Workshops on Computer Retrieval of Scientific Information 168
9.2 Seminars 172

TABLE OF CONTENTS (cont.)

Page

9.3 University Courses 173

9.4 Workbook for Modern Techniques in Chemical Information 176

9.5 Technical Presentations, Publications, and Mass Mailings 177

9.6 User Liaison 177

10. CENTER MANAGEMENT AND PROCEDURES 178

10.1 CSC Profile Handling Procedures 178

10.1.1 Receive Search Request 178

10.1.1.1 Review and Interpret Search Question 178

10.1.1.2 Conduct Manual Search 178

10.1.2 Profile Handling Procedures--New Profiles 179

10.1.3 Profile Handling Procedures--Modified Profiles 187

10.1.4 Designation of CSC Profile Numbers 187

10.1.4.1 CSC Profile Number 187

10.1.4.2 User-Type Classification 189

10.1.5 Keypunch Profile 189

10.1.6 Enter Profile in Input Data Deck 189

10.1.7 Check Output and Prepare Mailing 189

10.1.8 Monitoring Profiles 191

10.1.8.1 Monitoring New Profiles 191

10.1.8.2 Monitoring Existing Profiles 191

10.1.9 Dropping of Profiles 192

10.2 Center Files and User Records 192

10.2.1 User Record File 192

10.2.2 Corporate-User Record File 192

10.2.3 CSC User Profile History Book (Restricted Data) 192

10.2.4 CSC Profile Folder Files 194

10.2.5 Profile Correspondence File 194

10.2.6 Telephone Number File 194

TABLE OF CONTENTS (cont.)

Page

10.2.7 Profile Hit Evaluation File 194
10.2.8 Evaluation Report File 194
10.2.9 Abstract Number Card File 197
10.2.10 Profile Deck Modifications File 197
10.2.11 CSC Profile DROP Checklist 197
10.2.12 CSC Billing Forms 197
10.2.13 Subscription Card File 197
10.2.14 Profile Subject Index 199
10.2.15 Form Masters and Supplies 199
10.3 Tape Quality and Handling 202
10.3.1 Tape Quality--Physical Characteristics 202
10.3.2 Tape Quality--Readable, Mis-recorded Information 202
10.3.3 Tape Quality--Wrong Information 203
10.3.4 Tape Handling Procedures 203
10.4 Production Statistics and Cost 205
10.4.1 Computer Time per Program 205
10.4.2 CSC Time and Cost Summary Sheet 227
10.4.3 Production Run Summary--CSC 233
10.4.4 Profile Term Statistics 233
10.4.5 Profile Term Frequency Per Issue 238
10.4.6 Profile Term and Hit, Cost Data-Summaries 238
10.4.7 Personnel 251
10.5 Marketing 252
10.5.1 Mailings 252
10.5.2 Press Releases 253
10.5.3 Surveys 253
10.5.3.1 IEEE REFLECS Survey 253
10.5.3.2 Food Technology Survey 254
10.5.3.3 Market Survey 255
10.5.4 Pricing 266
10.5.5 Brochures 271
10.5.6 Contacts 271

TABLE OF CONTENTS (cont.)

Page

10.6 Contacts and Cooperative Arrangements 274

11. RESEARCH STATISTICS--COMPUTATIONAL LINGUISTICS AND ANALYTICAL STUDIES 276

  11.1 System Features 276

    11.1.1 Linking of Terms 276

    11.1.2 Truncation and Various Data Types 277

    11.1.3 Free Form Boolean Logic 283

    11.1.4 Weighting 283

  11.2 Terms--Profiles 288

  11.3 Hits (Output) 344

    11.3.1 Hits--Profiles 344

    11.3.2 Hits--Retrieved Citations 360

    11.3.3 Printed Hits 360

12. CONFERENCES, PRESENTATIONS, PUBLICATIONS AND PROFESSIONAL ACTIVITIES 386

13. REFERENCES 391

14. SUMMARY AND CONCLUSIONS 394

LIST OF TABLES

Table		Page
6-1	STATISTICS ON LENGTH, DATA FIELDS PER CITATION, AND KEYWORDS PER CITATION IN CA CONDENSATES	123
6-2	PERCENT OCCURRENCE OF DATA TYPES IN CA CITATIONS	125
6-3	AVERAGE LENGTH OF DATA ENTRIES BY DATA TYPE IN CA CITATIONS	126
6-4	CA TYPE:TOKEN RELATIONSHIPS	136
6-5	TYPE:TOKEN RATIOS	138
6-6	DATA BASE BIGRAM COMPARISON	145
8-1	RETURN OF EVALUATION FORMS VS. ISSUE	162
8-2	PRECISION VS. ISSUE	163
8-3	DISTRIBUTION OF AVERAGE PROFILE PRECISION	164
8-4	SUMMARY USER EVALUATION OF CA CONDENSATES CURRENT AWARENESS SERVICE	166
10-1	PERCENT AND RANGE PERCENT COMPUTER TIME PER PROGRAM	206
10-2	AVERAGE PERCENT COMPUTER TIME PER PROGRAM	206
10-3	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 71	207
10-4	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 72	208
10-5	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 73	209
10-6	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 74	210
10-7	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 75	211
10-8	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 76	212
10-9	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BIORI 71	213
10-10	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BA 51	214
10-11	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BA 52	215
10-12	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - EI 71	216
10-13	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - EI 72	217

LIST OF TABLES (cont.)

Table Page

10-14 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 71 240

10-15 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 72 241

10-16 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 73 242

10-17 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 74 243

10-18 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 75 244

10-19 PROFILE TERM, HIT, COST DATA VS. ISSUE - CA 76 245

10-20 PROFILE TERM, HIT, COST DATA VS. ISSUE - BIORI 71 246

10-21 PROFILE TERM, HIT, COST DATA VS. ISSUE - BA 51 247

10-22 PROFILE TERM, HIT, COST DATA VS. ISSUE - BA 52 248

10-23 PROFILE TERM, HIT, COST DATA VS. ISSUE - EI 71 249

10-24 PROFILE TERM, HIT, COST DATA VS. ISSUE - EI 72 250

10-25 EVALUATION OF SDI SYSTEM CHARACTERISTICS (PERCENT OF RESPONDENTS) 263

10-26 EVALUATION OF CHARACTERISTICS OF IITRI'S RETRIEVAL SYSTEM (PERCENT OF RESPONDENTS) 265

11-1 NUMBER OF TERMS OF VARIOUS TERM TYPES VS. TRUNCATION MODES USED 278

11-2 STATISTICAL OUTPUT FROM INPUTR 281

11-3 PERCENT OF PROFILES CONTAINING VARIOUS TERM TYPES VS. TRUNCATION MODE USED 282

11-4 NUMBER AND PERCENT OF PROFILES USING AND, OR, and NOT LOGIC VS. NUMBER OF TIMES EACH OPERATOR WAS USED IN A PROFILE 284

11-5 USE OF PARENTHETIC LOGIC IN PROFILES 285

11-6 USE OF NESTED LOGIC IN PROFILES 286

LIST OF FIGURES

Figure		Page
2-1	SDI SYSTEM GENERALIZED FLOW CHART	14
2-2	ENVIRONMENTS UNDER WHICH IITRI SOFTWARE HAS RUN	16
3-1	CA CONDENSATES OUTPUT--HEADER CARD	19
3-2	CA CONDENSATES OUTPUT--CITATION CARD	20
3-3	CA CONDENSATES OUTPUT--TRAILER CARD	21
3-4	BA PREVIEWS OUTPUT--HEADER CARD	22
3-5	BA PREVIEWS OUTPUT--CITATION CARD	23
3-6	BA PREVIEWS OUTPUT--TRAILER CARD	24
3-7	EI COMPENDEX OUTPUT--HEADER CARD	25
3-8	EI COMPENDEX OUTPUT--CITATION CARD	26
3-9	EI COMPENDEX OUTPUT--TRAILER CARD	27
4-1	PROFILE FORM--HEADER	32
4-2	PROFILE FORM--SEARCH TERMS	33
4-3	PROFILE FORM--LOGIC	34
4-4	TRUNCATION MODES	37
4-5	INFIX TRUNCATION	38
4-6	LEFT TRUNCATION	39
5-1	SDI SYSTEM GENERALIZED FLOW CHART	49
5-2	SDI SYSTEM DETAILED FLOW CHART. PART 1: DATA BASE INPUT	50
5-3	SDI SYSTEM DETAILED FLOW CHART. PART 2: PROFILE INPUT	51
5-4	SDI SYSTEM DETAILED FLOW CHART. PART 3: SEARCH	52
5-5	SDI SYSTEM DETAILED FLOW CHART. PART 4: OUTPUT	53
5-6	SDI SYSTEM DETAILED FLOW CHART. PART 5: STATISTICS	54
5-7	SDI SYSTEM DETAILED FLOW CHART. PART 6 (OPTIONAL): PRIVATE LIBRARIES SYSTEM EXTRACTION	55
5-8	DATA ELEMENTS AND IITRI DATA TYPE CODES	58
5-9	PREPROCESSOR SYSTEM	59
5-10	IITRI FORMATTED CITATION	60
6-1	NUMBER OF CITATIONS ON TAPE VS. ISSUE - CA 71-72	117

LIST OF FIGURES (cont.)

Figure Page

6-2 NUMBER OF CITATIONS ON TAPE VS. ISSUE - CA 73 118
6-3 NUMBER OF CITATIONS ON TAPE VS. ISSUE - CA 74 119
6-4 NUMBER OF CITATIONS ON TAPE VS. ISSUE - CA 75 120
6-5 NUMBER OF CITATIONS ON TAPE VS. ISSUE - CA 76 121
6-6 NUMBER OF CITATIONS ON TAPE VS. ISSUE - EI 71-72 122
6-7 CA ALPHABETICAL TERM FREQUENCY LIST 127
6-8 BA ALPHABETICAL TERM FREQUENCY LIST 128
6-9 EI ALPHABETICAL TERM FREQUENCY LIST 129
6-10 CA FREQUENCY-ORDERED TERM FREQUENCY LIST 130
6-11 BA FREQUENCY-ORDERED TERM FREQUENCY LIST 131
6-12 EI FREQUENCY-ORDERED TERM FREQUENCY LIST 132
6-13 TERM SETS WITH SIMILAR FREQUENCIES 135
6-14 REDUCTION IN TYPE:TOKEN RATIO AS A FUNCTION OF NUMBER OF TOKENS 137
6-15 CA KEY-LETTER-IN-CONTEXT INDEX 139
6-16 BA KEY-LETTER-IN-CONTEXT INDEX 140
6-17 EI KEY-LETTER-IN-CONTEXT INDEX 141
6-18 ALPHABETICAL BIGRAM LISTS 143
6-19 FREQUENCY-ORDERED BIGRAM LISTS 144
6-20 *ACID* APPEARANCES IN CA 147
6-21 *ACID* APPEARANCES IN EI 148
6-22 VARIATION IN TERM REPRESENTATION 149
7-1 SETS OF RELATED FRAGMENTS WITH UNCONTROLLED TRUNCATION 154
7-2 TRUNCATION GUIDE ENTRIES FOR THE CONCEPT "ANALYSIS" 156
8-1 USER EVALUATION REPORT FORM 161
9-1 WORKSHOP ANNOUNCEMENT BROCHURE--FRONT SIDE 169
9-2 WORKSHOP ANNOUNCEMENT BROCHURE--REVERSE SIDE 170
9-3 WORKSHOP PRECISION AND RECALL STATISTICS 171
9-4 AFFILIATION OF WORKSHOP ATTENDEES 173
10-1 CSC--SEARCH PROFILE-HEADER 180
10-2 CSC--SEARCH PROFILE-TERMS 181

LIST OF FIGURES (cont.)

Figure		Page
10-3	CSC--SEARCH PROFILE-LOGIC	182
10-4	PROFILE CHECK LIST	183
10-5	USER RECORD SHEET	184
10-6	CORPORATE-USER RECORD SHEET	185
10-7	CSC STANDARD "COVER" LETTER	186
10-8	PROFILE DECK MODIFICATION SHEET	188
10-9	CSC TRAILER CARD	190
10-10	DROP CHECKLIST	193
10-11	TELEPHONE NUMBER FILE	195
10-12	PROFILE HIT-EVALUATION FORM	196
10-13	BILLING FORM	198
10-14	SUBSCRIPTION CARD	200
10-15	PROFILE SUBJECT INDEX	201
10-16	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 72	218
10-17	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 73	219
10-18	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 74	220
10-19	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 75	221
10-20	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - CA 76	222
10-21	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BIORI 70-71	223
10-22	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BA 51	224
10-23	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - BA 52	225
10-24	PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE - EI 71-72	226
10-25	COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET - CA CONDENSATES	229
10-26	COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET - BA PREVIEWS	230
10-27	COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET - EI COMPENDEX	231

LIST OF FIGURES (cont.)

Figure		Page
10-28	TIME RECORD FROM CA PRODUCTION INPUTR RUN	232
10-29	PRODUCTION RUN SUMMARY--OVERALL SUMMARY	234
10-30	PRODUCTION RUN SUMMARY--PROFILE HIT DISTRIBUTION	235
10-31	PRODUCTION RUN SUMMARY--CORPORATE HIT DISTRIBUTION	236
10-32	PROFILE TERM STATISTICS	237
10-33	PROFILE TERM FREQUENCY PER ISSUE	239
10-34	FIELD INTERVIEW GUIDE--PAGE 1	256
10-35	FIELD INTERVIEW GUIDE--PAGE 2	257
10-36	FIELD INTERVIEW GUIDE--PAGE 3	258
10-37	FIELD INTERVIEW GUIDE--PAGE 4	259
10-38	FIELD INTERVIEW GUIDE--PAGE 5	260
10-39	FIELD INTERVIEW GUIDE--PAGE 6	261
10-40	CA PRICE SHEET	268
10-41	BA PRICE SHEET	269
10-42	EI PRICE SHEET	270
10-43	CSC BROCHURE--OUTSIDE	272
10-44	CSC BROCHURE--INSIDE	273
11-1	TERM AGGREGATION VS. ISSUE - CA 71-72	290
11-2	TERM AGGREGATION VS. ISSUE - CA 73	291
11-3	TERM AGGREGATION VS. ISSUE - CA 74	292
11-4	TERM AGGREGATION VS. ISSUE - CA 75	293
11-5	TERM AGGREGATION VS. ISSUE - CA 76	294
11-6	TERM AGGREGATION VS. ISSUE - BIORI 70-71	295
11-7	TERM AGGREGATION VS. ISSUE - BA 51	296
11-8	TERM AGGREGATION VS. ISSUE - BA 52	297
11-9	TERM AGGREGATION VS. ISSUE - EI 71-72	298
11-10	AGGREGATION REDUCTION RATIO VS. ISSUE - CA 71-72	299
11-11	AGGREGATION REDUCTION RATIO VS. ISSUE - CA 73	300
11-12	AGGREGATION REDUCTION RATIO VS. ISSUE - CA 74	301
11-13	AGGREGATION REDUCTION RATIO VS. ISSUE - CA 75	302

LIST OF FIGURES (cont.)

Figure		Page
11-14	AGGREGATION REDUCTION RATIO VS. ISSUE - CA 76	303
11-15	AGGREGATION REDUCTION RATIO VS. ISSUE - BIORI 70-71	304
11-16	AGGREGATION REDUCTION RATIO VS. ISSUE - BA 51	305
11-17	AGGREGATION REDUCTION RATIO VS. ISSUE - BA 52	306
11-18	AGGREGATION REDUCTION RATIO VS. ISSUE - EI 71-72	307
11-19	TERMS PER PROFILE VS. ISSUE - CA 71-72	308
11-20	TERMS PER PROFILE VS. ISSUE - CA 73	309
11-21	TERMS PER PROFILE VS. ISSUE - CA 74	310
11-22	TERMS PER PROFILE VS. ISSUE - CA 75	311
11-23	TERMS PER PROFILE VS. ISSUE - CA 76	312
11-24	TERMS PER PROFILE VS. ISSUE - BIORI 70-71	313
11-25	TERMS PER PROFILE VS. ISSUE - BA 51	314
11-26	TERMS PER PROFILE VS. ISSUE - BA 52	315
11-27	TERMS PER PROFILE VS. ISSUE - EI 71-72	316
11-28	COST PER TERM VS. ISSUE - CA 71-72	317
11-29	COST PER TERM VS. ISSUE - CA 73	318
11-30	COST PER TERM VS. ISSUE - CA 74	319
11-31	COST PER TERM VS. ISSUE - CA 75	320
11-32	COST PER TERM VS. ISSUE - CA 76	321
11-33	COST PER TERM VS. ISSUE - BIORI 70-71	322
11-34	COST PER TERM VS. ISSUE - BA 51	323
11-35	COST PER TERM VS. ISSUE - BA 52	324
11-36	COST PER TERM VS. ISSUE - EI 71-72	325
11-37	COST PER TERM PER CITATION VS. ISSUE - CA 71-72	326
11-38	COST PER TERM PER CITATION VS. ISSUE - CA 73	327
11-39	COST PER TERM PER CITATION VS. ISSUE - CA 74	328
11-40	COST PER TERM PER CITATION VS. ISSUE - CA 75	329

LIST OF FIGURES (cont.)

Figure | Page
11-41 COST PER TERM PER CITATION VS. ISSUE - CA 76 330
11-42 COST PER TERM PER CITATION VS. ISSUE - BIORI 70-71 331
11-43 COST PER TERM PER CITATION VS. ISSUE - BA 51 332
11-44 COST PER TERM PER CITATION VS. ISSUE - BA 52 333
11-45 COST PER TERM PER CITATION VS. ISSUE - EI 71-72 334
11-46 COST PER PROFILE VS. ISSUE - CA 71-72 335
11-47 COST PER PROFILE VS. ISSUE - CA 73 336
11-48 COST PER PROFILE VS. ISSUE - CA 74 337
11-49 COST PER PROFILE VS. ISSUE - CA 75 338
11-50 COST PER PROFILE VS. ISSUE - CA 76 339
11-51 COST PER PROFILE VS. ISSUE - BIORI 70-71 340
11-52 COST PER PROFILE VS. ISSUE - BA 51 341
11-53 COST PER PROFILE VS. ISSUE - BA 52 342
11-54 COST PER PROFILE VS. ISSUE - EI 72 343
11-55 HITS PER PROFILE VS. ISSUE - CA 71-72 345
11-56 HITS PER PROFILE VS. ISSUE - CA 73 346
11-57 HITS PER PROFILE VS. ISSUE - CA 74 347
11-58 HITS PER PROFILE VS. ISSUE - CA 75 348
11-59 HITS PER PROFILE VS. ISSUE - CA 76 349
11-60 HITS PER PROFILE VS. ISSUE - BIORI 70-71 350
11-61 HITS PER PROFILE VS. ISSUE - BA 51 351
11-62 HITS PER PROFILE VS. ISSUE - BA 52 352
11-63 HITS PER PROFILE VS. ISSUE - EI 71-72 353
11-64 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 71 354
11-65 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 72 355
11-66 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 73 356
11-67 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 74 357
11-68 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 75 358
11-69 NORMALIZED HITS PER PROFILE VS. ISSUE - CA 76 359

LIST OF FIGURES (cont.)

Figure		Page
11-70	HITS PER RETRIEVED CITATION VS. ISSUE - CA 72	362
11-71	HITS PER RETRIEVED CITATION VS. ISSUE - CA 73	363
11-72	HITS PER RETRIEVED CITATION VS. ISSUE - CA 74	364
11-73	HITS PER RETRIEVED CITATION VS. ISSUE - CA 75	365
11-74	HITS PER RETRIEVED CITATION VS. ISSUE - CA 76	366
11-75	HITS PER RETRIEVED CITATION VS. ISSUE - BIORI 70-71	367
11-76	HITS PER RETRIEVED CITATION VS. ISSUE - BA 51	368
11-77	HITS PER RETRIEVED CITATION VS. ISSUE - BA 52	369
11-78	HITS PER RETRIEVED CITATION VS. ISSUE - EI 71-72	370
11-79	NORMALIZED HITS VS. ISSUE - CA 71	371
11-80	NORMALIZED HITS VS. ISSUE - CA 72	372
11-81	NORMALIZED HITS VS. ISSUE - CA 73	373
11-82	NORMALIZED HITS VS. ISSUE - CA 74	374
11-83	NORMALIZED HITS VS. ISSUE - CA 75	375
11-84	NORMALIZED HITS VS. ISSUE - CA 76	376
11-85	COST PER HIT VS. ISSUE - CA 71-72	377
11-86	COST PER HIT VS. ISSUE - CA 73	378
11-87	COST PER HIT VS. ISSUE - CA 74	379
11-88	COST PER HIT VS. ISSUE - CA 75	380
11-89	COST PER HIT VS. ISSUE - CA 76	381
11-90	COST PER HIT VS. ISSUE - BIORI 70-71	382
11-91	COST PER HIT VS. ISSUE - BA 51	383
11-92	COST PER HIT VS. ISSUE - BA 52	384
11-93	COST PER HIT VS. ISSUE - EI 71-72	385

FOUR-YEAR SUMMARY

EDUCATIONAL AND COMMERCIAL UTILIZATION OF A CHEMICAL INFORMATION CENTER

1. INTRODUCTION

1.1 Report Description

This report summarizes and organizes all of the significant findings and information recorded in the previous IITRI reports C6156-1 through C6156-17. The Summary Report provides a comprehensive overview of the research performed in establishing and operating the Computer Search Center (CSC) at IITRI, and provides an overall analysis of these data. The evolution of ideas and design parameters is also given in historical perspective. Thus this report can be used to trace the course of the project in lieu of piecing together the Quarterly Reports that detailed work in progress.

The Summary Report is composed of fourteen major sections that detail the research activities from conception and design through implementation and operation. The second part of this section, the INTRODUCTION, provides the history and background of the project. It presents the perspective from which to view the balance of the report.

Section 2 covers the COMPUTER SEARCH CENTER DESIGN AND DEVELOPMENT. Our initial objectives, initial system design made to meet those objectives, and development of the design are discussed. The bases for our original decisions on hardware, programming language and program features for installation independence are given.

Section 3 is concerned with the SERVICES provided by the CSC. These include Selective Dissemination of Information (SDI) and retrospective searches, the Private Libraries System and software installation. Sections 4 and 5 elaborate on these services, with 4 covering PROFILE PREPARATION AND MODIFICATION, including discussions of profile forms, formats and features, while 5 describes the SOFTWARE SYSTEM. Section 5 contains both descriptive information on topics such as the IITRI file structure, data base format conversion, search strategy and logic evaluation, and definitive information on the program set, core requirements, and files.

Sections 6, 7, 8, and 9 relate to the relationships among the users, the data bases and CSC. Section 6 gives DATA BASE CHARACTERISTICS AND COMPARISONS. The next section describes the USER AIDS we have developed: Search Manual and Supplemental Guides, KLIC Indexes, Term Frequency lists and Truncation Guide. The next section in this group covers USER EVALUATION AND FEEDBACK. Section 9 covers EDUCATION-USER LIAISON by discussing workshops, seminars, courses, and the Workbook on Modern Techniques for Chemical Information.

Section 10 describes all the functions of CENTER MANAGEMENT AND PROCEDURES necessary to serve the users. It includes such topics as profile, data base, and user record handling as well as internal statistics, marketing, and relationships with other centers.

Section 11 covers the RESEARCH activities of the program. Many analytical studies of data bases and linguistic analyses of profile-citation interfaces were made in the course of this project. Some basic facts were discovered, such as finding that lexicographical ordering based on letters from left to right in a word is a poor ordering form upon which to base a search algorithm. The applications of this and other findings to text searching are discussed.

The last three sections, 12, 13, and 14 present listings of conferences, presentations, publications, and professional activities carried out in conjunction with the program; a list of REFERENCES and our SUMMARY AND CONCLUSIONS.

1.2 History and Background

The proliferation of chemical literature over the past several decades has been a growing source of concern to both the professional scientist and his management. There are now over 300,000 papers per year referenced in Chemical Abstracts, and 250,000 per year in Biological Abstracts. Several years ago a government research executive was quoted as saying: "If the research program cost $100,000 or less, it is less expensive to do it again than to make sure it has not been done before!" This statement, fortunately, is no longer true. Many of the principal secondary sources--indexing and abstracting journals as well as other collections of information--have been established for searching these new data bases to provide scientists and engineers with an inexpensive means of coping with the scientific literature. Currently more than two million scientific and technical papers are published each year and even with the use of abstracting and indexing journals, it is no longer feasible for the average scientist to keep up in his own field if he must rely on manual searching.

Numerous solutions to the information explosion problem have been posed, such as reducing the number of articles published, publishing only summaries or abstracts of articles, or retaining full documentation on magnetic tape only, and announcing the existence of the information to persons in the appropriate subject areas. The implementation of such solutions in our "publish or perish" society where publications effect both salary and ego boosting would seem to indicate that printed publications, either full articles or as shortened versions, are here to stay. Hence, the machine-readable versions of these or their surrogates will need to be searched by information centers.

The cost of keyboarding or otherwise preparing large machine-readable files is high, and until recent years when the preparation of machine-readable records was done for purposes of computerized typesetting or to speed up publication, the cost of inputting information could seldom be justified on the basis of information retrieval. Computer-readable files are now being produced in significant numbers and even though, in most cases, the file is created as a by-product of publication activity, the file does exist and can be searched.

A survey by the American Institute of Physics[1] identified 50 commercially available scientific and technical data bases. The Directory of Computerized Information in Science and Technology[2] has identified several hundred additional data bases--most of which are specialized and small. There are currently perhaps 10-20 popular data bases and many more that enjoy limited use. The Association of Scientific Information Dissemination Centers (ASIDIC)[3], Cooperative Data Management Committee, recently published the ASIDIC Survey of Information Center Services and found that the 56 responding centers identified 48 publically-available data bases that they are processing either for SDI (selective dissemination of information) or retrospective searches.

In the late 1960's the Office of Science Information Services (OSIS) of the National Science Foundation (NSF) recognized the need for data base services and research and development regarding the data bases, data base services, and operational aspects of centers that handle machine-readable data bases. Accordingly, NSF provided seed money for several "university based information centers". These are located at the University of Pittsburg, the University of Georgia, Lehigh University, the University of California (UCLA) and at IIT Research Institute (IITRI). Although IITRI is a not-for-profit contract research organization and not a university, it is affiliated with Illinois Institute of Technology (IIT).

The IITRI Computer Search Center (CSC) was established in 1968 and was designed as a one-stop information center to meet user needs by providing a variety of desired sources and services with minimal restrictions and a high degree of flexibility. Services include both current awareness (SDI) and retrospective searches tailored to a user's or organization's needs. Users of the Center are scientists and engineers in industry, universities, and other research organizations.

The SDI system has been operational since September 1969 and CSC offers services from Chemical Abstracts' Condensates (CA), Biological Abstracts' Previews (BA), and Engineering Index's COMPENDEX (EI) on a production basis. CSC plans to add the International Food Information Service data base in the fall of 1972.

2. COMPUTER SEARCH CENTER DESIGN AND DEVELOPMENT

2.1 Objectives

Chemists generate, need, and use chemical information as indicated by the existence of a large number of primary chemical journals, by the size and growth rate of the secondary abstracting journals, and by the existence of chemical libraries in many commercial, educational, and government research and development installations. The more than 100,000 chemists in the American Chemical Society spend a significant amount of time perusing the literature. It was noted in an article in the July 28, 1969 issue of Chemical and Engineering News that the average amount of time spent by an industrial chemist on current awareness reading is 7.5 hours per week.

ACCESS, the listing of journals by Chemical Abstracts Service (CAS), names more than 20,000 chemical or chemistry-related journals. This number does not take into account in-house publications or research and development reports prepared by industry, government, and government contractors both within and outside of the United States.

Traditionally, the rehandling and distribution of technical information has been done by means of printed publications. Because the volume of scientific literature has grown so large, it has become necessary to employ automation and new techniques to make the information available to users within a reasonable time span. Much has been done and reported regarding computer techniques for composition, storage, search, and retrieval of chemical information. However, it is necessary to utilize the newer techniques and sources and to train the users--the bench chemists--who are familiar with the standard and traditional sources and means of obtaining information in the use of the new technology.

There is a large volume of chemical information that now exists in machine-readable form and there are many chemists who are the potential users of this information. A potential market exists but there is a real problem in devising methods of bringing the users to the new information sources or disseminating the information from these sources to users. Information scientists at IITRI are helping to solve this problem by the operation of the Computer Search Center.

2.2 Design and Development

The CSC system was designed to provide a variety of information storage and retrieval-type services from a multiplicity of existing and future data bases, with numerous profile options, flexible search strategies, variable sort options, and varying output media. This was to be done in a manner that would permit us to use one generalized software system that would be easy to modify and alter and would be machine independent and installation independent.

The general objectives led to the establishment of design requirements and the development of special features for the CSC system. Requirements included: program transferability; machine independence and installation independence; ability to handle numerous data bases; development of general purpose programs; and modularity. Special features included: aggregation of profile terms; left and right truncation of terms; free-form Boolean logic; removal of redundant search terms; options for sorting of output; options for media on which output is printed; and designation of hit terms, index terms, and weight on each output citation form.

Because none of the computer search programs available at the time met all of the criteria required by the Center, and because of the need to handle a variety of data bases, new general purpose computer programs were written. The compiler language PL/1 was employed to achieve machine and installation independence and hence a high degree of program transferability.

CSC programs were initially written and debugged using the RUSH (Remote-User-Shared-Hardware) interactive programming system. Using a terminal at IITRI, programs were written, compiled, and debugged on a 360/50 in Palo Alto, California. RUSH is a dialect of PL/1 and programs were developed avoiding those features and statements in RUSH that were not currently in PL/1. Once the programs were written and debugged on RUSH, they were converted to PL/1 and run on several 360's in the Chicago area. The transition from RUSH to PL/1 went very smoothly.

The programs were written in a modular fashion so that changes, additions, and deletions could be readily accommodated. A separate block was written for each separate operation within a program. The basic functions provided by the programs are source tape format conversion, profile preparation, search, output generation, and maintenance of statistics. The programs are described in detail in Section 5 of this report.

The basic set of programs was written, tested, and put in production in September 1969. At that time a pilot group of users prepared 146 profiles for searches of CA Condensates. Subsequently, BA Previews and COMPENDEX were added to the production system.

The number of users and profiles has varied from time to time as new users have been brought into the system and experimental profiles were tried. Users represented industry, academia, and government with the majority being from industry.

Throughout the course of the project and as production data accumulated we have made continuing efforts to update, streamline, and increase the effectiveness of our computer programs. These efforts have been rewarded as is evidenced by a very great reduction in computer processing time required for the weekly production searches (see Section 10).

In addition to the creation of an operational computer search, retrieval, and dissemination system, IITRI has instituted educational and training programs, the purpose being not only to develop a center, but to ensure its continuing use in the future. This objective led to the development of a Search Manual for profile preparation, the development of a workbook in Modern Techniques in Chemical Information, the teaching of a new academic course at Illinois Institute of Technology, and the presentation of seminars. A detailed discussion of the educational aspects of the project is given in Section 9 of this report.

2.3 Programming Language Selection

The CSC design criteria of software transferability, machine independence, and installation independence together with the desire to carry out coding tasks in a relatively short time period while generating modular, flexible general purpose programs led to the decision to develop software in a higher level compiler language rather than in a machine language or assembly language.

We investigated several compiler languages such as FORTRAN, COBOL, ALGOL, and PL/1 and selected PL/1 because of its flexibility, generality, power, and modularity. The program goal of producing programs that could be transferred to other organizations ruled out the machine dependent machine level and assembler level languages. Of the higher level compiler languages, PL/1 appeared to offer the best balance of flexibility, generality, power, and modularity necessary for the other goals of generating a program set that could change as data bases changed, incorporate new data bases and contain the features desired by users.

PL/1 is currently available on IBM 360 and 370 series hardware, which comprises the majority of computer installations. Burroughs has announced a PL/1 compiler and one for the Digital Equipment Corporation PDP-10 is nearly ready. Univac, CDC, and others are preparing PL/1 compilers. Thus, PL/1 will shortly be quite machine independent. Even considering the 360-70 family as a limitation of sorts, there is wide variation among the many models in this series. CSC programs have run on over 15 different configurations of 360's and 370's with no problems. If currently not machine independent, PL/1 assuredly has a high degree of configuration-independence.

Many of PL/1's features are eminently suitable for text processing. These include character and bit string handling functions, structure variables, hierarchical data structures and arrays, list processing capabilities, and device independent I/O. By using these features, we have been able to implement all of the system concepts we have evolved in the PL/1 language. In no case did the language limit our design options. This is due to the rich syntax of the language and speaks very highly for its flexibility.

PL/1 is admittedly not as efficient as an assembler level language and there are usually many ways to do any operation--with varying degrees of efficiency. However, the use of modular programming techniques and the power of the language have overcome this lower execution efficiency. Since we were able to try out design modifications in very short amounts of time and without disrupting a production schedule, we were able to devote less time to programming and coding and more time to investigation of what really goes on in a bibliographic search system. This enabled us to test six different search techniques and to develop such concepts as that of the Least Common Bigram which more than offset the efficiency differences of PL/1 and assembler level languages. Such multiple testing would not have been possible within reasonable constraints of time, dollars, and the realities of a production activity without a compiler level language such as PL/1. Modular programming techniques, easily implemented in PL/1, allowed us to make changes in portions of the set (to accomodate a new data base, for example, or to react to a data base format change) with no interruption to production and without changing all the programs in the set.

In addition, PL/1 is quickly learned and it is possible to familiarize new staff members with the overall programming system in a relatively short time. With a sophisticated set of assembler language programs the termination of a staff member is likely to be a more traumatic experience than is the case with PL/1.

2.4 Computer

One of CSC's objectives was the development of programs that could run at a variety of installations. Inasmuch as the IBM 360 family of computers represents a large segment of the computer field and PL/1 compilers are available, we decided to program for the 360-70 series computers. Initially, the choice of PL/1 tied us to 360-70 machines but since more than 50 percent of the computers in the country fall in that category this limitation did not pose a serious constraint. Subsequently, Burroughs has announced a PL/1 compiler, one is under development for Digital Equipment Corporation's PDP-10, and proprietary compilers exist for CDC, Honeywell, and Univac equipment so the boundaries seem to be relaxing. Although, for instance, FORTRAN compilers are available for many makes of computers, transferability is not a surety. Only parts of FORTRAN as a whole are basic to all the hardware, and thus we would have imposed quite severe limitations upon ourselves with that choice.

CSC programs will run on IBM 360's from a Model 40 on up. They require a minimum of two tape drives, one or more disks, and, assuming approximately 3000 search terms (200 profiles of 15 terms each), 256K bytes of core storage.

We believe that our design philosophy has been a serviceable one. We have demonstrated the utility of PL/1 and use of the IBM 360 in that we were able to develop a sophisticated information retrieval system and get into a production mode in a relatively short period of time. We have been able to test many alternative programming approaches and implement changes to the system as needed and we have run the system on 15 different computers. (See Section 2.6).

2.5 Modularity and Program Modules

Programs were developed in a modular fashion in order to permit changes, additions, replacements, and deletions in programs and program modules without affecting the entire system. A separate block was written for each separate operation within a program. There are five basic functions carried out by the programs. The programs together with the names of the eleven specific program modules that accomplish the functions are:

	Program Function	Program Module Name
(1)	Preparation of data base input	DBCOPY (Data Base Copy) FORCON (Format Conversion) IFCOPY (IITRI-Format Copy)
(2)	Preparation of profile input	DKEDIT (Deck Edit) MINIPUP (Mini Profile Update Program) INPUTR (Profile Input Preparation Routine)
(3)	Search data base for profiles	SEARCH
(4)	Preparation of search output	HITTER (Hit Recorder) DBCARD (Data Base Card Format) DBOCP (Data Base Output Control Program)
(5)	Statistics generation	STIXA (Statistics)

A twelfth program which is optional is call PLSXT (Private Libraries System Extraction) and is used for extracting data from the SDI system to be used as input for the Private Libraries System (PLS). PLS is a software system for creating and maintaining private files or subset data bases. It is discussed in Section 5.8.

The interrelationships of the programs and component modules can be seen in the simplified flow chart Figure 2-1. Details regarding the specific programs and their relationships to each other and to the files that are used for communication between programs and modules are given in Sections 5.3 and 5.5.

Via the modularity feature the total software system is constructed of multiple individually replaceable and changeable building blocks. Individual modules or programs can be changed or replaced without affecting other portions of the same program or other programs (and specific subroutines can be called for in certain cases and not others) thus permitting a high degree of flexibility.

An example of this feature can be seen in the fact that the format conversion module (FORCON) is different for each data base yet the programs and files it interfaces with are unaffected. Also the output card formatting module (DBCARD) is different for each data base depending on which of the data elements contained on the data base are to be displayed on the output cards. DBCARD interfaces with other portions of the system which remain the same regardless of whether the specific DBCARD program is for Chemical Abstracts (CACARD), Biological Abstracts (BACARD), or Engineering Index (EICARD).

In addition to this replacement feature is the ability to revise specific programs as needed. For example, if a data base supplier adds a new data element to his files or changes format, we can change the FORCON program to accomodate the supplier change. This can be done readily and easily. In fact, we have made hundreds of minor changes to individual program modules and have never interrupted the production activity of our weekly runs. More significantly, we have been able to make major changes to programs and conduct comparative tests quickly and inexpensively. For example, the basic search strategy has been changed several times and other approaches have been tested. These tests are discussed in Section 5.6.

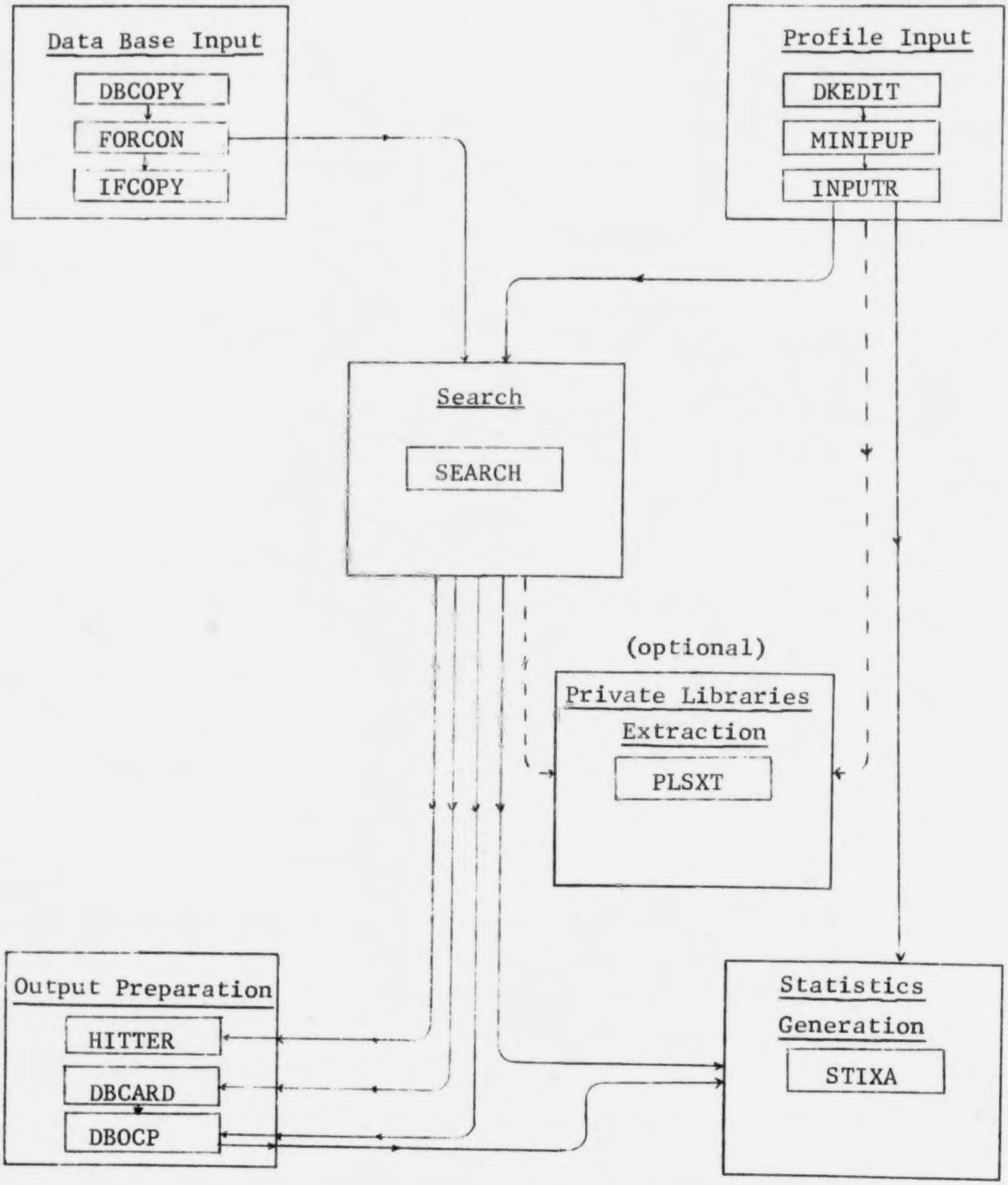

Data Base Input
DBCOPY
FORCON
IFCOPY
Profile Input
DKEDIT
MINIPUP
INPUTR
Search
SEARCH
(optional)
Private Libraries Extraction
PLSXT
Output Preparation
HITTER
DBCARD
DBOCP
Statistics Generation
STIXA

Figure 2-1
SDI SYSTEM GENERALIZED FLOW CHART

2.6 Transferability--Machine and Installation Independence

Machine and installation independence permit transferability of software, which was one of the CSC design goals. Reasons for this design goal were: anticipation (realized within a year) of a hardware change at IITRI; the desire to install our software in organizations that needed an internal SDI system; and the desire to conduct profile writing workshops and training courses both on-site and at other locations. Successful achievement of this design goal is evident from the fact that we have installed the system at several industrial organizations and have run the programs at 15 different computer facilities with no real difficulties. Preparation of appropriate JCL is usually all that is required. Figure 2-2 indicates the variety of hardware, processors, versions of the operating system and releases of the PL/1 compiler that we have used.

Hardware:	IBM 360	Models:	40 50 65 67 75
	IBM 370	Model:	155
	Any computer with PL/1 Compiler		
Processors:			MFT MVT PCP HASP
Operating System Versions:			15-16 17 18 19 19.6 20 21
PL/1 Compiler Releases:			4.1 5 5.2

Figure 2-2

ENVIRONMENTS UNDER WHICH IITRI SOFTWARE HAS RUN

3. SERVICES

The CSC was designed to provide a variety of services. Among those currently offered are SDI (Selective Dissemination of Information), retrospective searches conducted either by computer or manually, private library development and maintenance, and software installation. SDI is the principal service offered by CSC.

3.1 SDI

The current awareness or SDI (Selective Dissemination of Information) system has been operational since September 1969, and the Computer Search Center (CSC) is now offering services from Chemical Abstracts Condensates, Biological Abstracts, Bioresearch Index and Engineering Index's COMPENDEX. Searches of other data bases will be added depending on user needs.

The SDI system was designed to include many user-oriented features, including: full free form Boolean logic with any degree of nesting; many searchable elements; all forms of term truncation; weighting; sort options; and print media options.

One may include searchable elements as positive or negative search terms, i.e., one may require the presence or absence of any particular search term to qualify a citation as a "hit" citation. Among the searchable elements are:

Subject terms appearing in titles, text, or as index terms

Author names

Company names

Journal names as represented by the standard ASTM CODEN

Country

CA section numbers

BA CROSS Codes

BA BIOSYSTEMATIC Codes

EI Card-A-Lert Codes

The search terms may be single words, multi-word terms, phrases, or portions of words.

Output may be sorted according to user preference by author, weight, or citation number. Standard output is prepared on 5" x 8" cards. Provisions can be made for printing output on paper or multilith masters for further reproduction and dissemination within an organization.

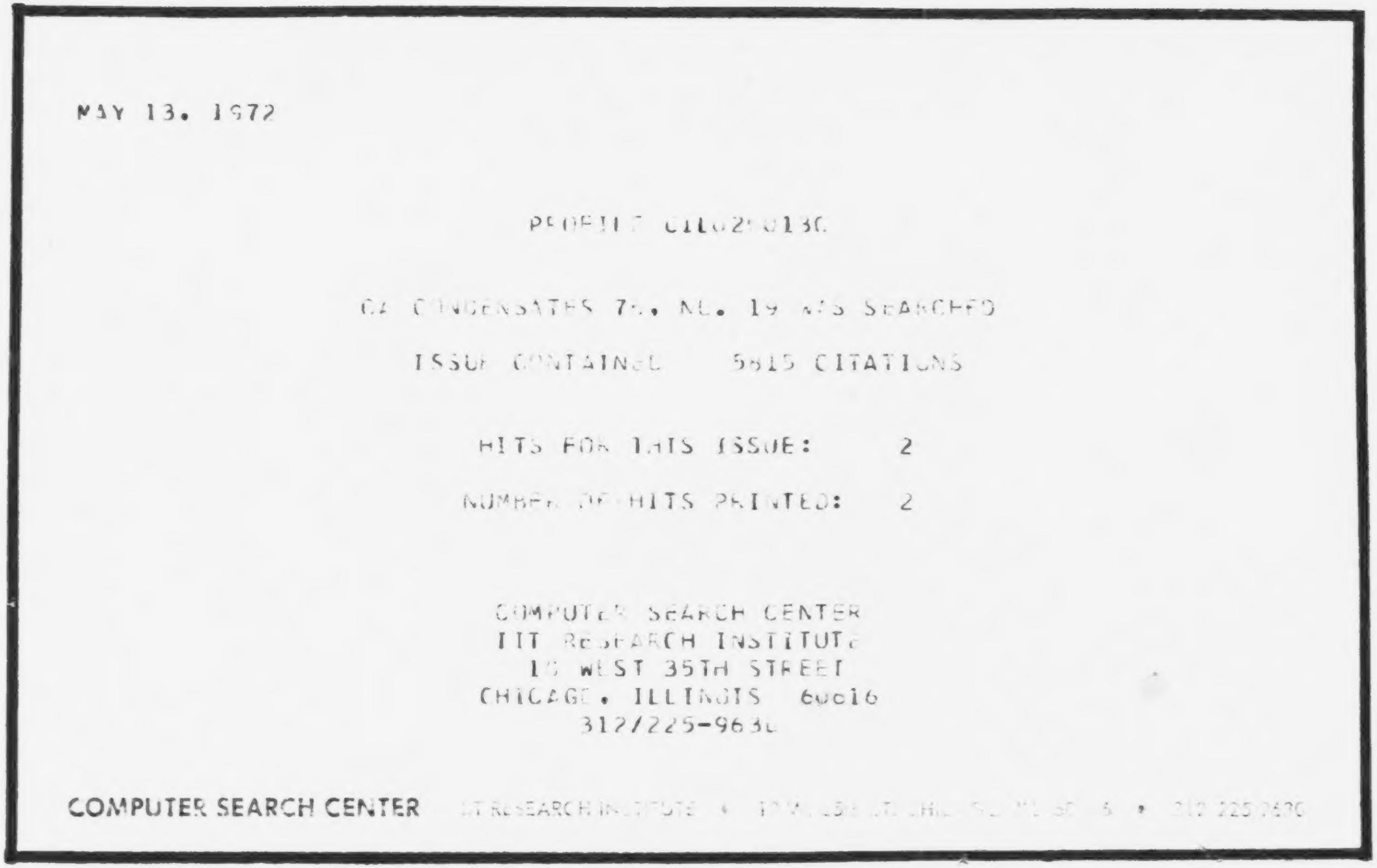
The standard output sent to users is printed on three types of cards--header, citation, and trailer. The header card as shown indicates: the user profile number, the tape service and issue of the tape that was searched, the number of citations that were on tape, the number of citations that were hit citations for the user's profile, the number of citations that were printed, and the date of the search. Examples of header cards for CA, BA, and EI are shown in Figures 3-1, 3-4, and 3-7.

There is one 5" x 8" citation card for each hit. A citation card includes: citation number; tape source including volume and issue number; profile number; authors (as many as are given on the source tape) and corporate authors; full title; primary source information including journal volume, issue, date, pages, and CODEN; index terms; abstracts; codes and any other significant information that may have been included on the source tape; search terms present, i.e., those profile terms that were hit terms for the particular citation; and weight for the citation. Examples of citation cards with the data items that are specific to a given data base are shown for CA, BA, and EI in Figures 3-2, 3-5 and 3-8. Trailer cards listing the total citations in a user's output are shown in figures 3-3, 3-6, and 3-9.

Searches are conducted and output sent to users weekly, biweekly, or monthly in accordance with the frequency of the particular data base to be searched.

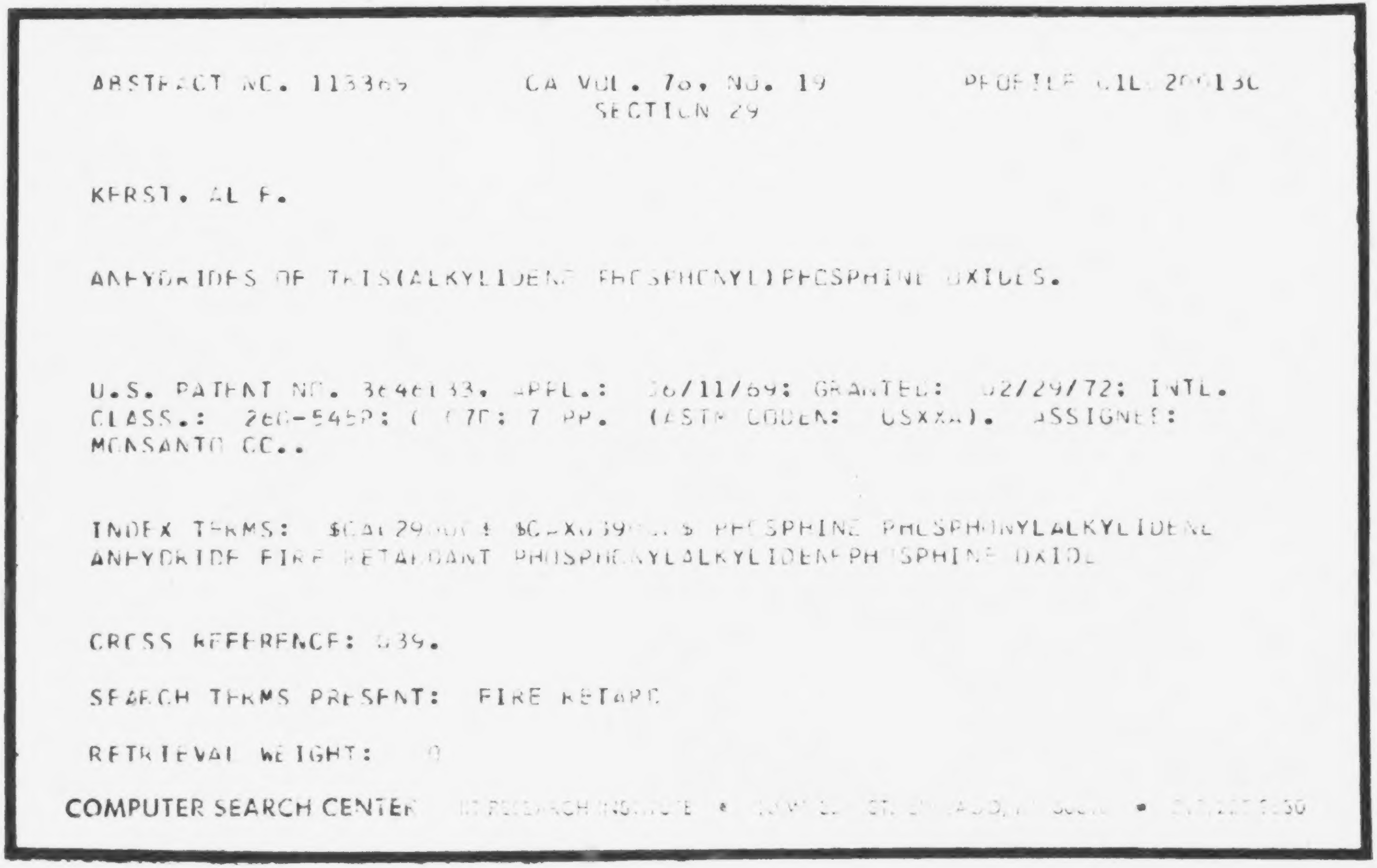
3.2 Retrospective Searches

Retrospective searches, either manual or by machine, are provided on request. The price is dependent on the number



```
MAY 13. 1972

                    PROFILE C1L020130

          CA CONDENSATES 76, NO. 19 WAS SEARCHED

            ISSUE CONTAINED     5815 CITATIONS

                HITS FOR THIS ISSUE:      2

              NUMBER OF HITS PRINTED:     2


                COMPUTER SEARCH CENTER
                IIT RESEARCH INSTITUTE
                 10 WEST 35TH STREET
               CHICAGO, ILLINOIS  60616
                     312/225-9630

COMPUTER SEARCH CENTER   IIT RESEARCH INSTITUTE • 10 W. 35TH ST. CHICAGO, ILLINOIS 60616 • 312 225-9630
```

Figure 3-1
CA CONDENSATES OUTPUT-HEADER CARD

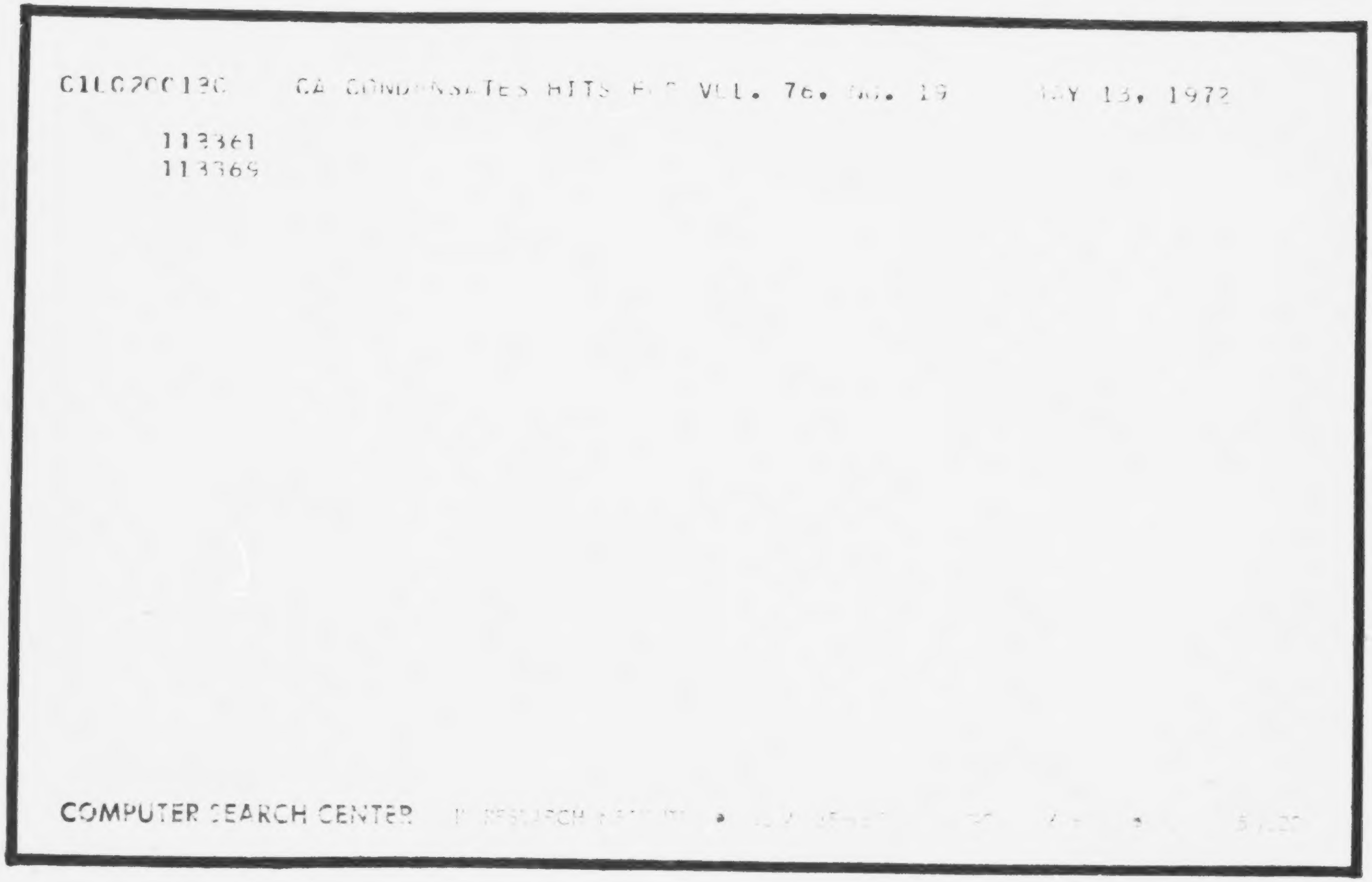


```
ABSTRACT NO. 113365        CA VOL. 76, NO. 19        PROFILE CIL 200130
                                SECTION 29


KERST, AL F.


ANHYDRIDES OF TRIS(ALKYLIDENE PHOSPHONYL)PHOSPHINE OXIDES.



U.S. PATENT NO. 3646133. APPL.:  06/11/69; GRANTED:  02/29/72; INTL.
CLASS.:  260-545P; C 07F; 7 PP.  (ASTM CODEN:  USXXAM).  ASSIGNEE:
MONSANTO CO..


INDEX TERMS:  $CA129[illegible] $C-X039[illegible] PHOSPHINE PHOSPHONYLALKYLIDENE
ANHYDRIDE FIRE RETARDANT PHOSPHONYLALKYLIDENEPHOSPHINE OXIDE


CROSS REFERENCE: 039.

SEARCH TERMS PRESENT:  FIRE RETARD

RETRIEVAL WEIGHT:   0
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • [illegible] • [illegible]

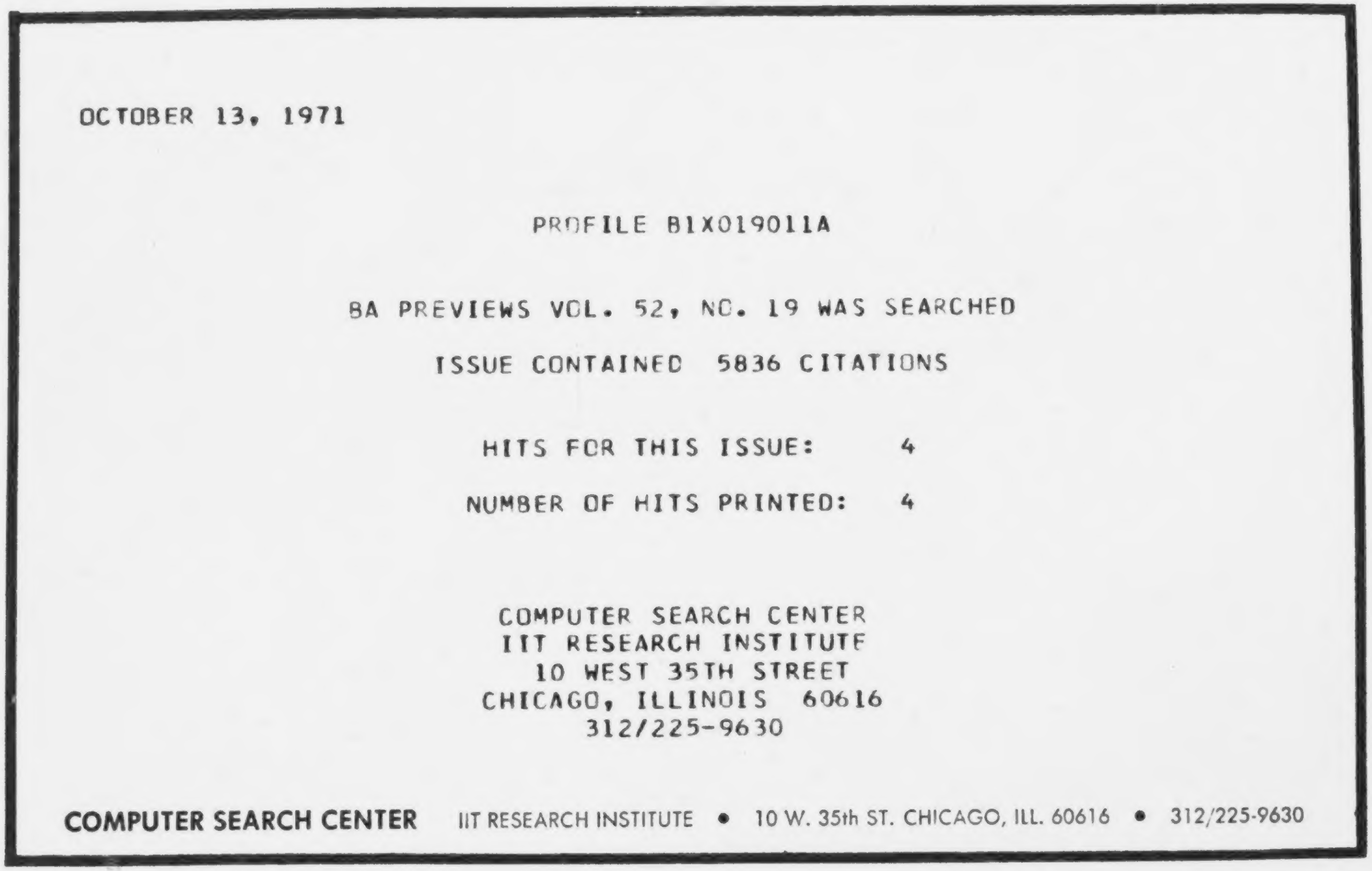
Figure 3-2

CA CONDENSATES OUTPUT-CITATION CARD



```
C1LC2CC13C     CA CONDENSATES HITS FOR VOL. 76, NO. 19      MAY 13, 1972

     113361
     113369





COMPUTER SEARCH CENTER   [illegible] RESEARCH [illegible] • [illegible] • [illegible]
```

Figure 3-3
CA CONDENSATES OUTPUT-TRAILER CARD

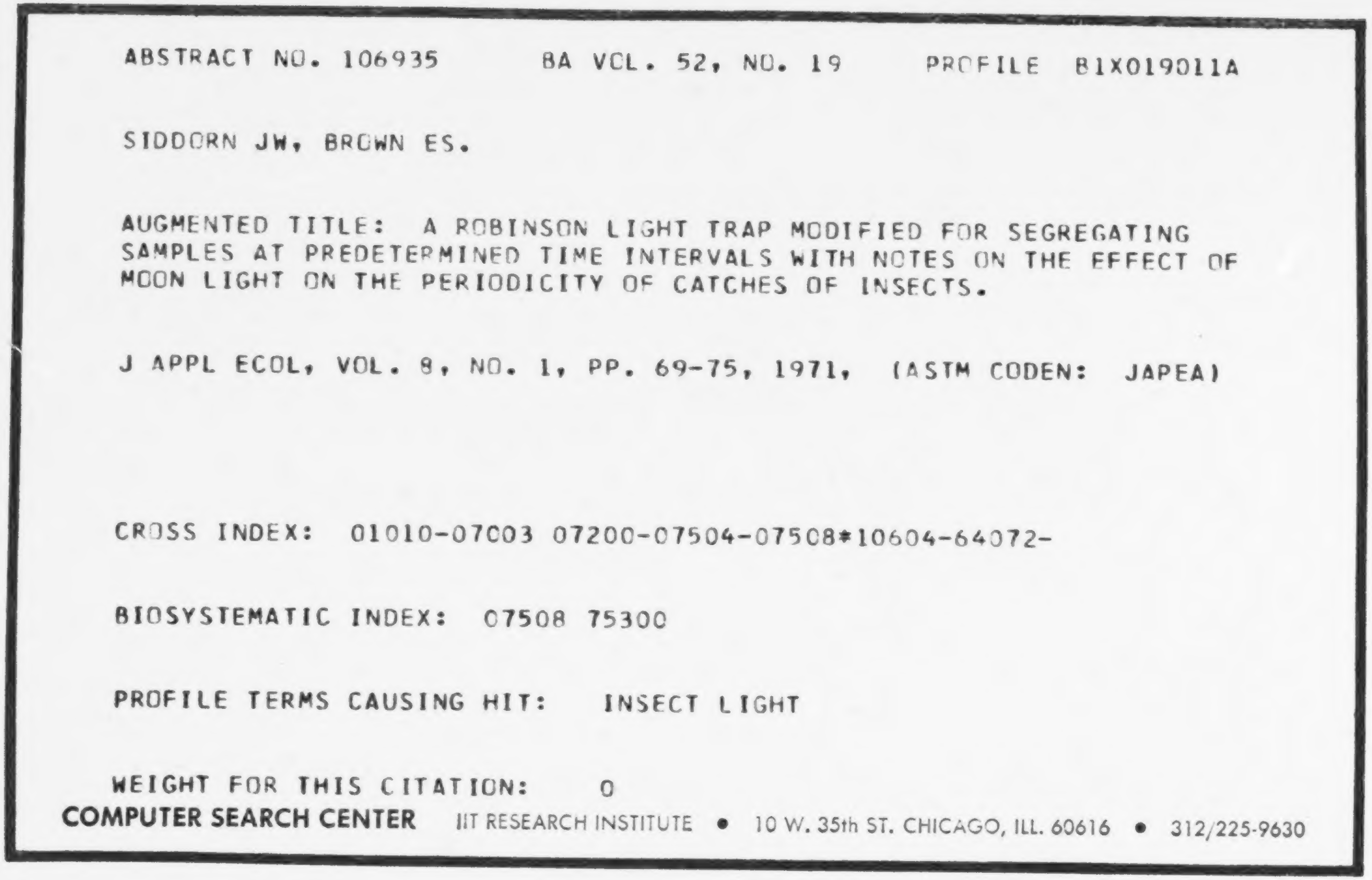


```
OCTOBER 13, 1971

                    PROFILE B1X019011A

          BA PREVIEWS VOL. 52, NO. 19 WAS SEARCHED

             ISSUE CONTAINED  5836 CITATIONS

               HITS FOR THIS ISSUE:      4

              NUMBER OF HITS PRINTED:    4


                 COMPUTER SEARCH CENTER
                 IIT RESEARCH INSTITUTE
                  10 WEST 35TH STREET
                CHICAGO, ILLINOIS  60616
                     312/225-9630
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

Figure 3-4

BA PREVIEWS OUTPUT-HEADER CARD

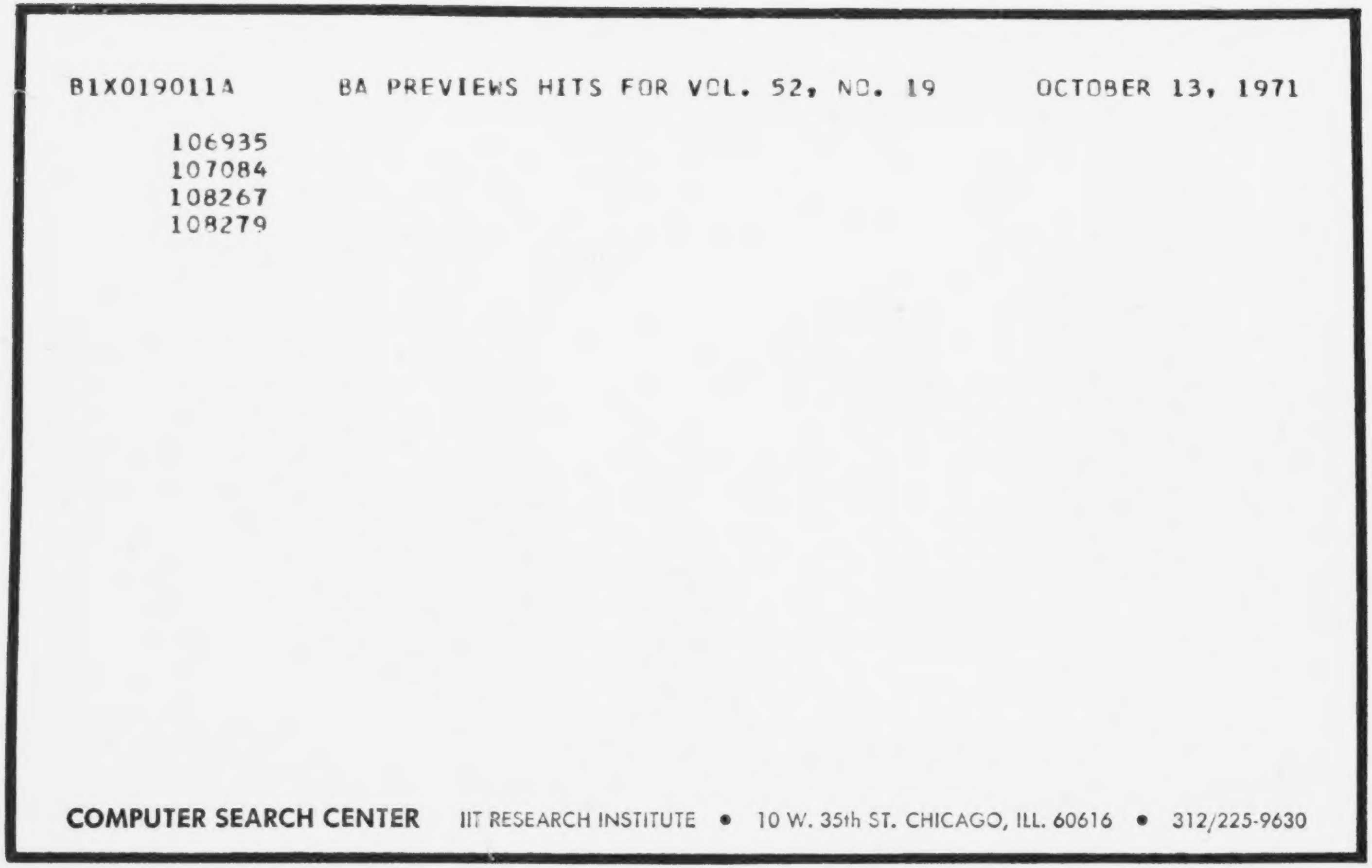


```
ABSTRACT NO. 106935      BA VOL. 52, NO. 19     PROFILE  B1X019011A

SIDDORN JW, BROWN ES.

AUGMENTED TITLE:  A ROBINSON LIGHT TRAP MODIFIED FOR SEGREGATING
SAMPLES AT PREDETERMINED TIME INTERVALS WITH NOTES ON THE EFFECT OF
MOON LIGHT ON THE PERIODICITY OF CATCHES OF INSECTS.

J APPL ECOL, VOL. 8, NO. 1, PP. 69-75, 1971,  (ASTM CODEN:  JAPEA)



CROSS INDEX:  01010-07003 07200-07504-07508*10604-64072-

BIOSYSTEMATIC INDEX:  07508 75300

PROFILE TERMS CAUSING HIT:   INSECT LIGHT

WEIGHT FOR THIS CITATION:    0
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

Figure 3-5

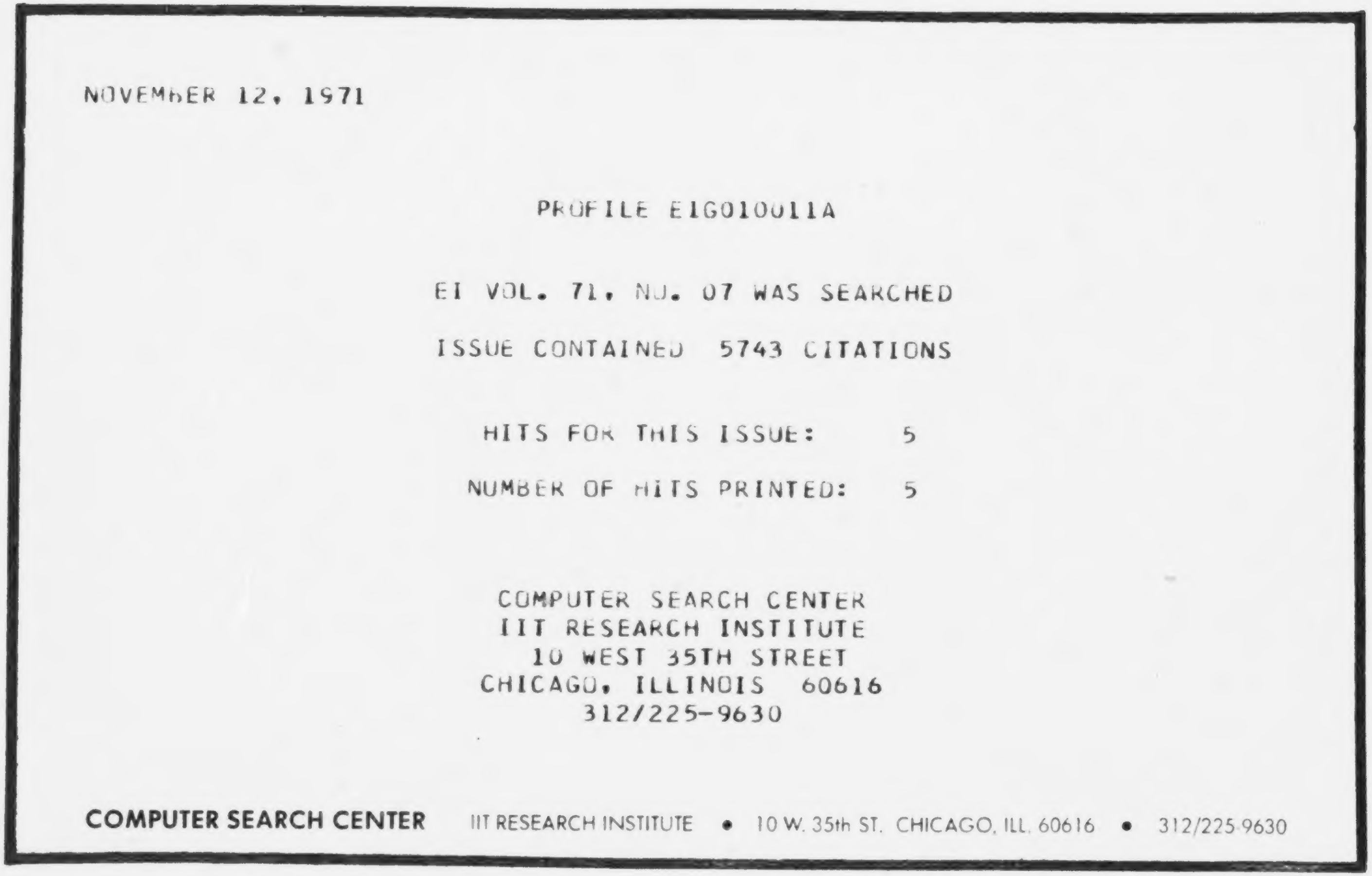
BA PREVIEWS OUTPUT-CITATION CARD



```
B1X019011A       BA PREVIEWS HITS FOR VOL. 52, NO. 19       OCTOBER 13, 1971

         106935
         107084
         108267
         108279
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

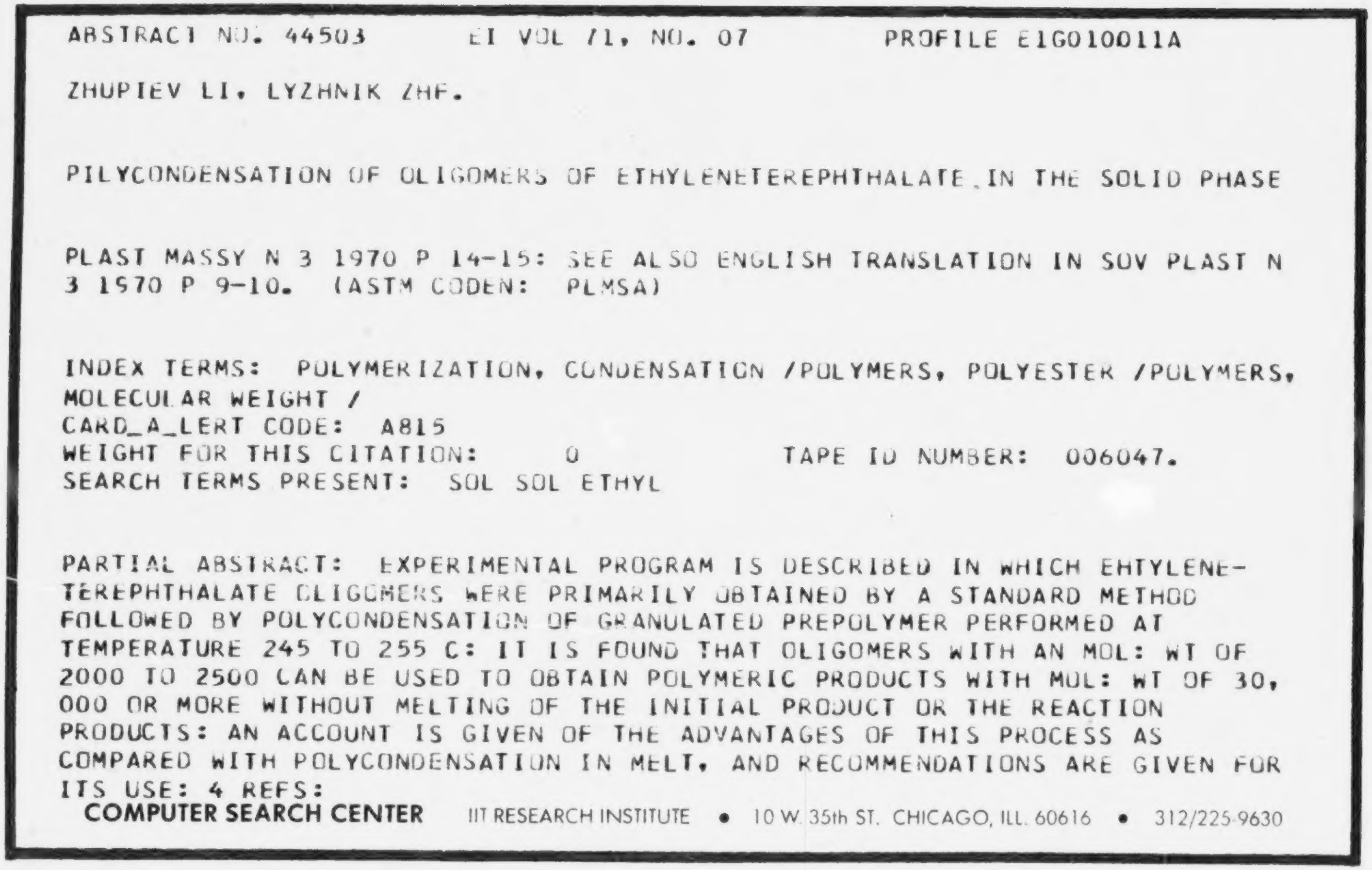
Figure 3-6
BA PREVIEWS OUTPUT-TRAILER CARD



```
NOVEMBER 12, 1971

                    PROFILE EIG010011A

              EI VOL. 71, NO. 07 WAS SEARCHED

              ISSUE CONTAINED  5743 CITATIONS

                HITS FOR THIS ISSUE:     5

               NUMBER OF HITS PRINTED:   5


                  COMPUTER SEARCH CENTER
                  IIT RESEARCH INSTITUTE
                   10 WEST 35TH STREET
                CHICAGO, ILLINOIS  60616
                      312/225-9630
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

Figure 3-7
EI COMPENDEX OUTPUT-HEADER CARD

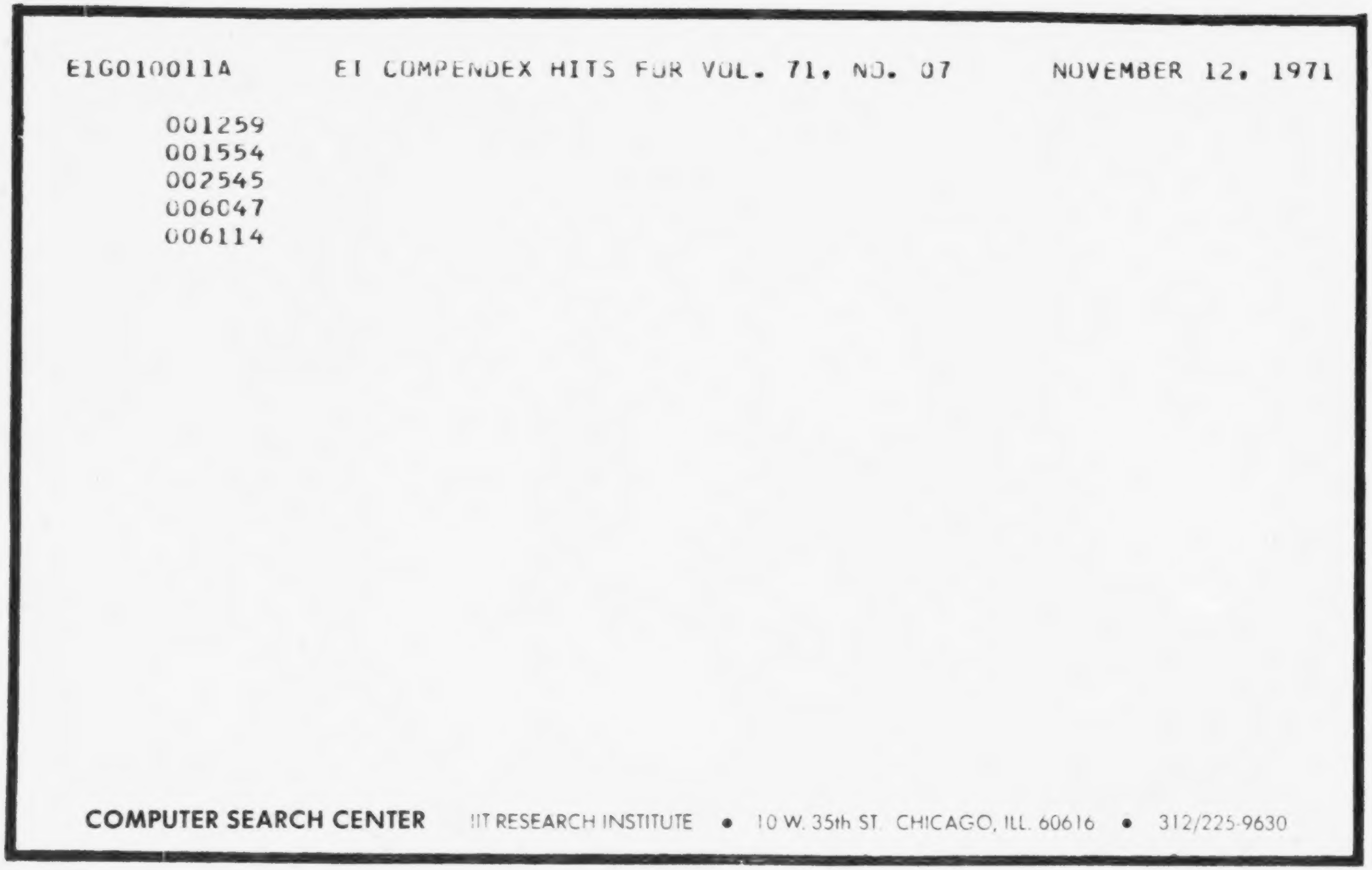


```
ABSTRACT NO. 44503      EI VOL 71, NO. 07        PROFILE E1G010011A

ZHUPIEV LI, LYZHNIK ZHF.

PILYCONDENSATION OF OLIGOMERS OF ETHYLENETEREPHTHALATE.IN THE SOLID PHASE

PLAST MASSY N 3 1970 P 14-15: SEE ALSO ENGLISH TRANSLATION IN SOV PLAST N
3 1970 P 9-10.  (ASTM CODEN:  PLMSA)

INDEX TERMS:  POLYMERIZATION, CONDENSATION /POLYMERS, POLYESTER /POLYMERS,
MOLECULAR WEIGHT /
CARD_A_LERT CODE:  A815
WEIGHT FOR THIS CITATION:      0           TAPE ID NUMBER:  006047.
SEARCH TERMS PRESENT:  SOL SOL ETHYL

PARTIAL ABSTRACT:  EXPERIMENTAL PROGRAM IS DESCRIBED IN WHICH EHTYLENE-
TEREPHTHALATE OLIGOMERS WERE PRIMARILY OBTAINED BY A STANDARD METHOD
FOLLOWED BY POLYCONDENSATION OF GRANULATED PREPOLYMER PERFORMED AT
TEMPERATURE 245 TO 255 C: IT IS FOUND THAT OLIGOMERS WITH AN MOL: WT OF
2000 TO 2500 CAN BE USED TO OBTAIN POLYMERIC PRODUCTS WITH MOL: WT OF 30,
000 OR MORE WITHOUT MELTING OF THE INITIAL PRODUCT OR THE REACTION
PRODUCTS: AN ACCOUNT IS GIVEN OF THE ADVANTAGES OF THIS PROCESS AS
COMPARED WITH POLYCONDENSATION IN MELT, AND RECOMMENDATIONS ARE GIVEN FOR
ITS USE: 4 REFS:
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

Figure 3-8

EI COMPENDEX OUTPUT-CITATION CARD



```
EIG010011A      EI COMPENDEX HITS FOR VOL. 71, NO. 07      NOVEMBER 12, 1971

      001259
      001554
      002545
      006047
      006114
```

COMPUTER SEARCH CENTER IIT RESEARCH INSTITUTE • 10 W. 35th ST. CHICAGO, ILL. 60616 • 312/225-9630

Figure 3-9
EI COMPENDEX OUTPUT-TRAILER CARD

of years searched, the size of the data base (or portion of a data base), the number of search terms, and the frequency of search terms in the data base. We are currently developing programs to provide retrospective searches of indexes, inverted files, and/or merged data bases and are planning a service to search the forthcoming CAS Integrated Subject File.

In all cases a judgment is made as to whether a machine search or manual search would be most effective and efficient, and a recommendation and a cost estimate are then given to the requestor. A single term search, for example, can certainly be carried out more efficiently by manually searching indexes, whereas a search that employs numerous search terms and/or complicated logic might best be done by machine.

3.3 Private Libraries

Through the Private Libraries System we can create tailor-made machine-readable data bases from document collections, company report files, and other information resources specified by a client. Each such data base, while specifically designed in terms of content to reflect the particular subject material in the information collection, is represented in uniform format on tape. The IITRI data base format allows specification of the types of data elements, such as author, keyword, or report number within the record itself. Different numbers of elements and different elements can be specified for individual records and/or data bases. The length of each element is also variable and not predetermined. The flexibility of this format allows us to generate data bases from widely varying types of information. Yet, our software works with any data base in this format. We have programs that allow addition, deletion, and modification of entire records or parts of records in order to update, modify, and improve the data at whatever time the client wishes. Also, bibliographies, concordances, etc., can be inexpensively produced from the data base. A private library that is specially-tailored for a client is maintained for the client and searched exclusively by client organization personnel.

3.4 Software Installation

IITRI will install its software at a user's installation, providing complete checkout of the software and training of operational personnel. The installation service includes:

- Program source decks and complete documentation, including flowcharts and narrative comments
- Installation and program checkout on-site, including JCL and data set preparation
- Training in running the system, including error recovery
- Detailed training, including test run experience, in profile construction and refinement
- IITRI's unique user aids
- On-site production run test under IITRI supervision
- Maintenance and development support of software
- Consultation service for user problems

Installation of IITRI's software includes many services beyond the handing over of an operational set of programs. The software itself, of course, is the sine qua non of the installation. We reproduce a full set of source decks, then compile them and run a complete test run with the decks that will be turned over to the user along with complete documentation. These decks are then taken to the user's computer facility and checked out by an IITRI specialist. At this time JCL for the system is made up, disk and tape files assigned, and the software checked out on the user's machine. Basic instruction in running the system is given to the personnel who will be actively involved in the production use of the system, and a test run performed including doctored data designed to cause specific errors--both to demonstrate typical malfunctions and to test the operational staff's ability to correct errors and proceed.

The key phase of the procedure, however, is not the software installation, but the profile construction training which is performed in a special workshop at IITRI. At this point the profile coordinators and users are instructed in the techniques necessary to produce effective profiles. We supply a complete set of our user aids and detailed instruction in their use. Several test runs are made to allow the user's staff to get a first-hand knowledge of the techniques of profile construction and refinement. IITRI's unique combination of experience and capability are available to the user throughout the set-up period and thereafter, in the interest of providing the user with the ability to produce effective profiles.

The final phase of the installation is an on-site production test under IITRI supervision. Two complete production runs are done in one week, with every phase of the operation carefully checked and monitored by IITRI personnel. At this point the user's staff should demonstrate an ability to run the system and recover from errors caused by the sorts of faulty input that occurs in normal production.

After the installation is complete, IITRI's fund of experience and detailed knowledge of the internal logic of the system is available to the user by telephone or mail. In addition, program improvements will be provided as they are introduced for a period agreed on. Thereafter future improvements will be available for a limited charge, allowing the user to keep up with the steady improvement in system efficiency and effectiveness that results from IITRI's continuing investment in refining and optimizing existing programs and developing better methods.

Thus our installation service is a complete package of training and operational components. The unique combination provides the user with a comprehensive system which he is capable of using maximally.

4. PROFILE PREPARATION AND MODIFICATION

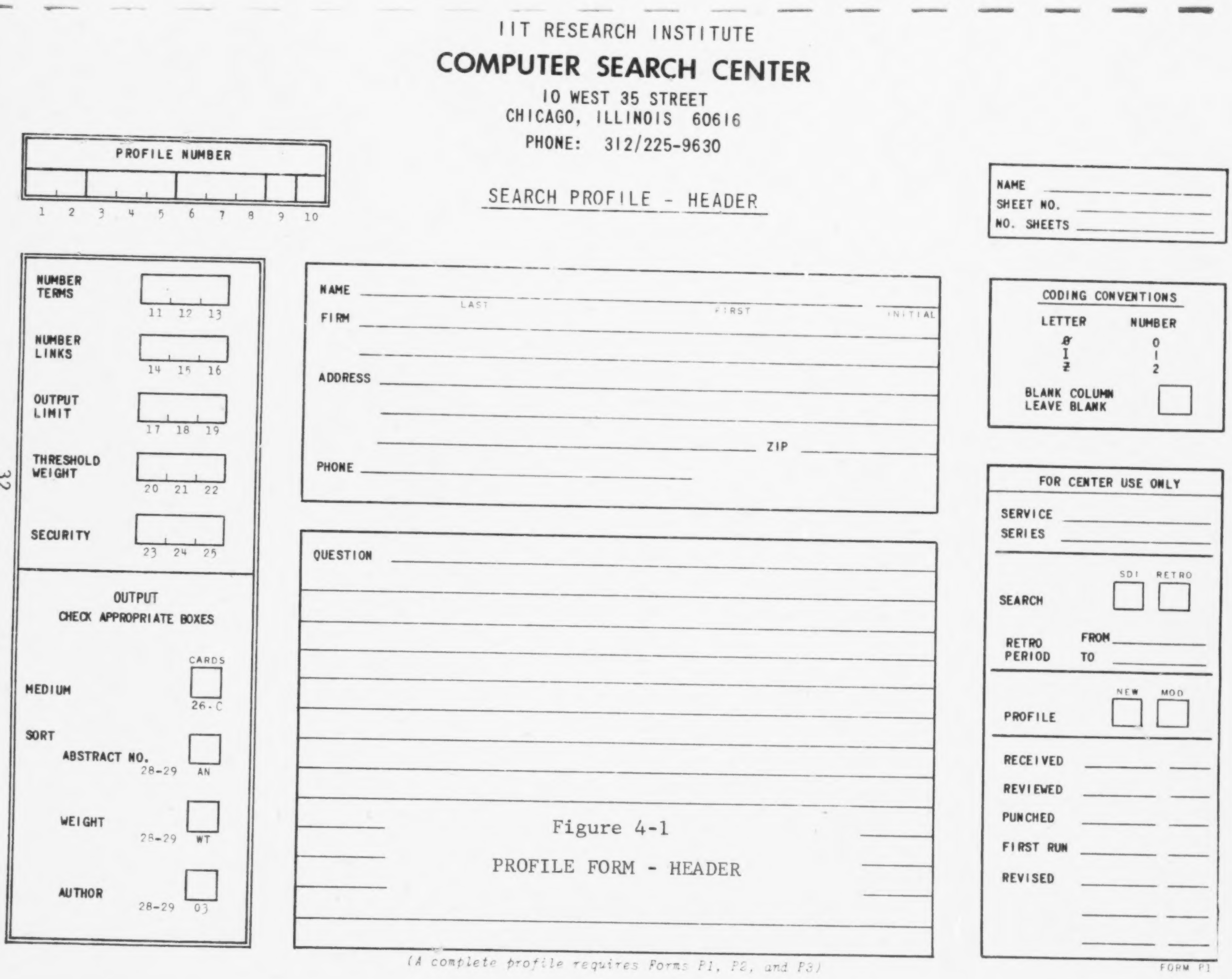
4.1 Profile Forms

The search profile is the primary input into the system. It is a representation of a question by a user in the terminology of a data base and coded according to the conventions of the search system. Search terms are the data elements constituting a search profile and are common to the terminology of both the search question and the data file. Profile information, user identification, and the search question are entered on the Header profile form illustrated in Figure 4-1. All search terms relevant to a particular search profile are listed on the Terms coding form shown in Figure 4-2. Each term is assigned a referent (term number) in the sequence by which the terms are listed on the coding form. Truncation mode and term type are also entered on the Term coding form.

Terms that are semantically associated can be linked together in a single expression. Linked terms are synonyms, related terms, or hierarchical (broader, narrower) terms. A link designator represents the associated terms and can be used to simplify the logic expression and to facilitate the cumulation of weights. The link designator, a single character from the set A-Z, is entered on the coding form.

Weights are numerical values assigned to search terms that indicate their relative significance to the user. The weights augment the logic of the expression and increase the retrieval effectiveness of a profile. Term weights can range from 0 to 9. If the weight option is chosen, the output can be sorted in weighted order, with the highest weighted items printed first. A print cutoff can be designated by the user to eliminate printing of the lowest weighted items.

Two modes of weighting are used to circumvent the problem arising from the presence of synonyms or related terms in a logic expression. A noncumulative mode selects in a link only the weight of the highest weighted term that is found in a



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

SEARCH PROFILE - HEADER

PROFILE NUMBER

1 2 3 4 5 6 7 8 9 10

NUMBER TERMS	11 12 13
NUMBER LINKS	14 15 16
OUTPUT LIMIT	17 18 19
THRESHOLD WEIGHT	20 21 22
SECURITY	23 24 25

OUTPUT
CHECK APPROPRIATE BOXES

MEDIUM CARDS ☐ 26-C

SORT

ABSTRACT NO.	28-29	☐	AN
WEIGHT	28-29	☐	WT
AUTHOR	28-29	☐	03

NAME ______ LAST ______ FIRST ______ INITIAL
FIRM ______
ADDRESS ______
______ **ZIP** ______
PHONE ______

QUESTION ______

NAME ______
SHEET NO. ______
NO. SHEETS ______

CODING CONVENTIONS

LETTER	**NUMBER**
Ø	0
I	1
Ƶ	2

BLANK COLUMN LEAVE BLANK ☐

FOR CENTER USE ONLY

SERVICE ______
SERIES ______

SEARCH SDI ☐ RETRO ☐

RETRO PERIOD FROM ______ TO ______

PROFILE NEW ☐ MOD ☐

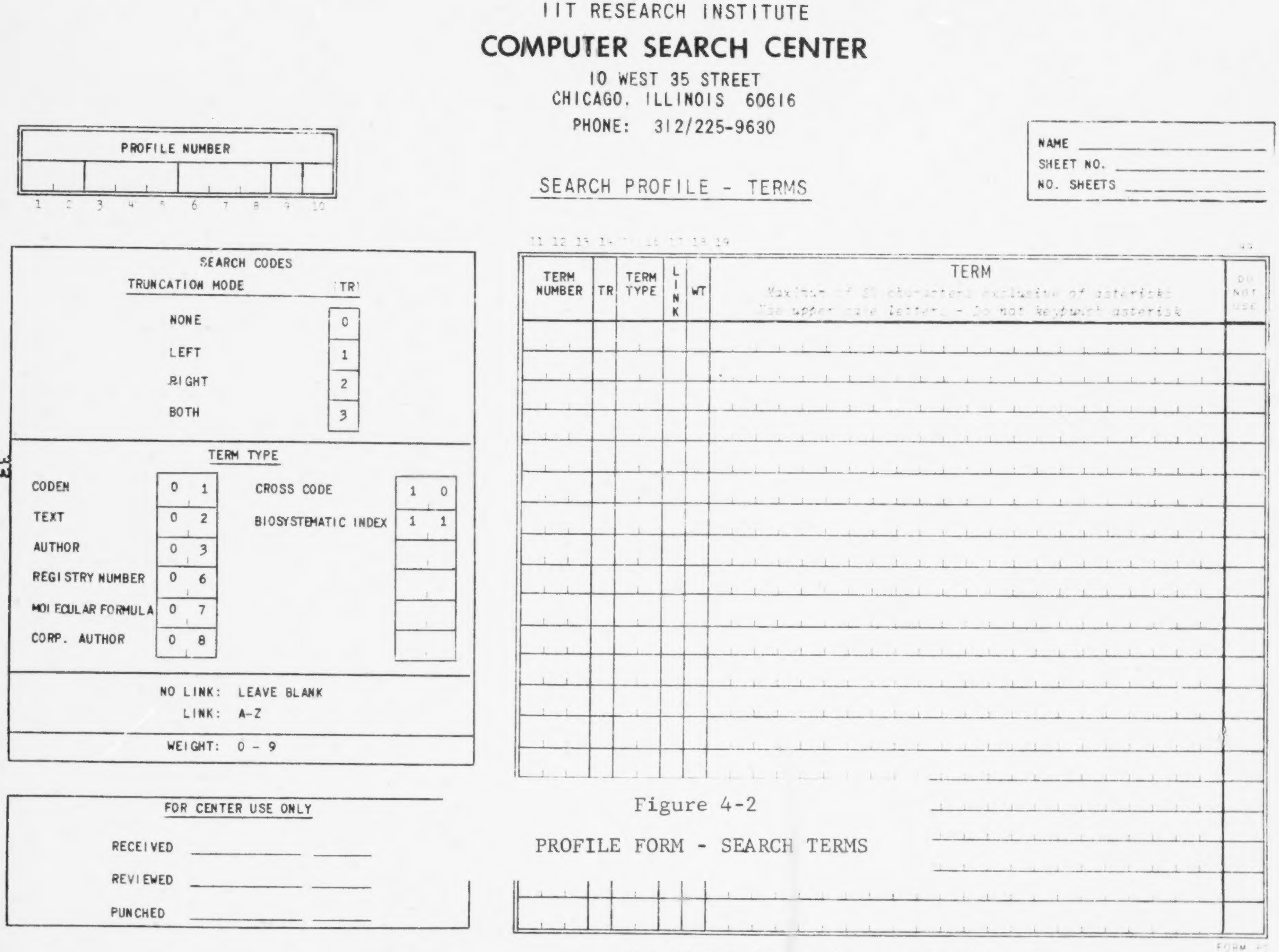
RECEIVED ______ ______
REVIEWED ______ ______
PUNCHED ______ ______
FIRST RUN ______ ______
REVISED ______ ______

Figure 4-1

PROFILE FORM - HEADER

(A complete profile requires Forms P1, P2, and P3)

FORM P1



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

SEARCH PROFILE - TERMS

PROFILE NUMBER

1 2 3 4 5 6 7 8 9 10

NAME ______
SHEET NO. ______
NO. SHEETS ______

SEARCH CODES

TRUNCATION MODE (TR)

NONE	0
LEFT	1
RIGHT	2
BOTH	3

TERM TYPE

CODEN	0 1	CROSS CODE	1 0
TEXT	0 2	BIOSYSTEMATIC INDEX	1 1
AUTHOR	0 3		
REGISTRY NUMBER	0 6		
MOLECULAR FORMULA	0 7		
CORP. AUTHOR	0 8		

NO LINK: LEAVE BLANK
LINK: A-Z

WEIGHT: 0 - 9

FOR CENTER USE ONLY

RECEIVED ______ ______
REVIEWED ______ ______
PUNCHED ______ ______

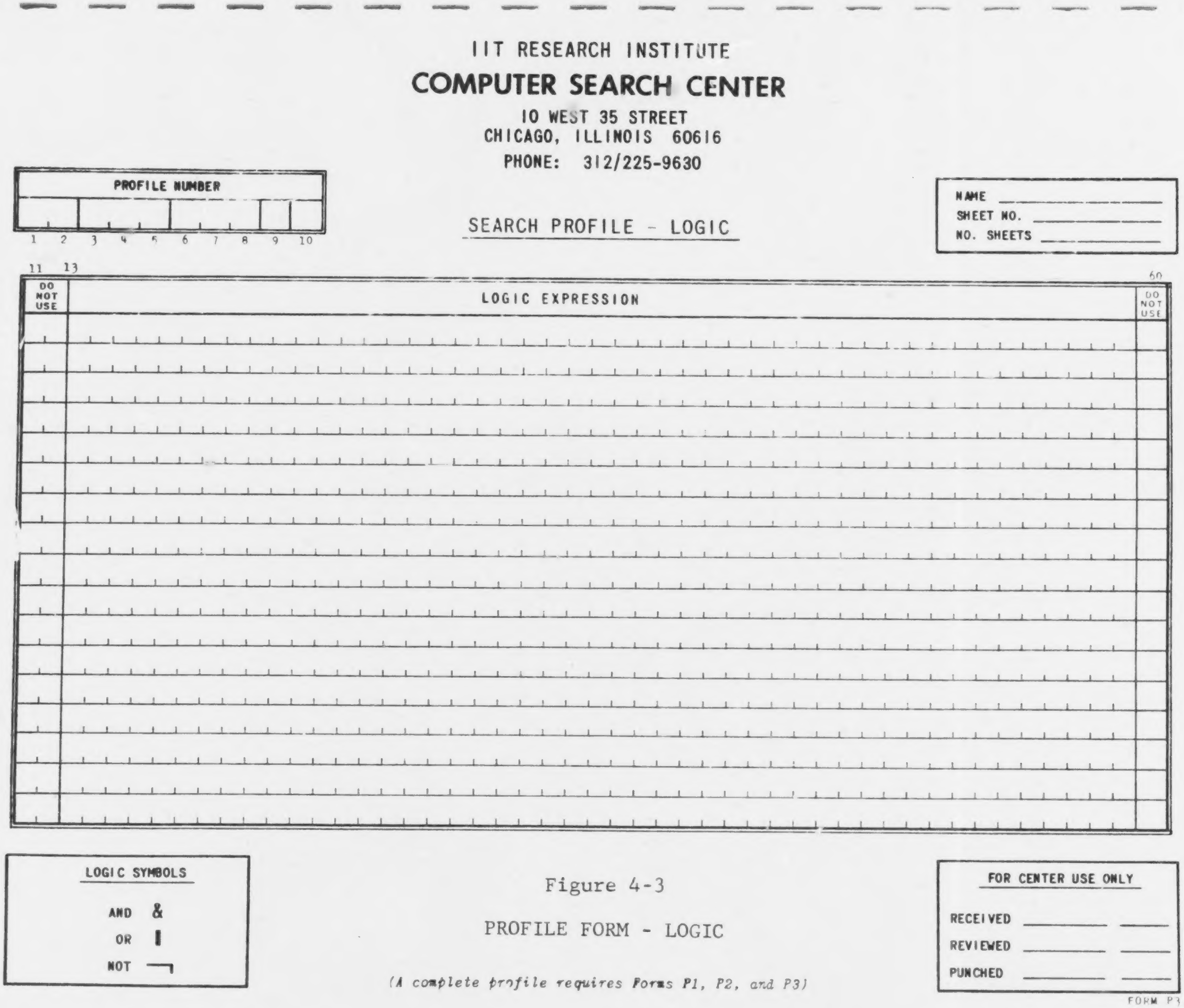
11 12 13 14 15 16 17 18 19

TERM NUMBER	TR	TERM TYPE	LINK	WT	TERM Maximum of [illegible] characters exclusive of asterisks Use upper case letters - Do not keypunch asterisk	DO NOT USE

(A complete profile requires Forms P1, P2, and P3)

Figure 4-2

PROFILE FORM - SEARCH TERMS



IIT RESEARCH INSTITUTE
COMPUTER SEARCH CENTER
10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

PROFILE NUMBER
1 2 3 4 5 6 7 8 9 10

SEARCH PROFILE - LOGIC

NAME ____________
SHEET NO. ____________
NO. SHEETS ____________

11 13 60

DO NOT USE	LOGIC EXPRESSION	DO NOT USE

LOGIC SYMBOLS
AND &
OR |
NOT ¬

FOR CENTER USE ONLY
RECEIVED ________ ______
REVIEWED ________ ______
PUNCHED ________ ______

FORM P3

Figure 4-3

PROFILE FORM - LOGIC

(A complete profile requires Forms P1, P2, and P3)

retrieved citation. A cumulative mode adds the weights of all other terms. A threshold weight can be specified and only citations that satisfy the logic expression and whose weight is equal to or greater than the specified threshold are retrieved.

The terms and links are associated in a logic expression on the Logic coding form illustrated in Figure 4-3. The logical operators AND, OR, and NOT can be written in any free form Boolean expression with any level of nesting.

4.2 Profile Options and Features

The principal features built into the system to achieve effective profiles and to allow wide flexibility in the way terms can be used are the following: wide variety of term types; all forms of term truncation; full free form Boolean logic with any degree of nesting to relate terms to each other; grouping or linking of similar terms; and weighting of terms according to user assignment of relevance. Statistics regarding the use of various profile options are given in Section 11.

4.2.1 Terms

One may include searchable elements as positive or negative search terms, i.e., one may require the presence or absence of any particular search term to qualify a citation as a "hit". The following are term options available to a user.

- Terms--anything other than single character
  - Single word
  - Multi-word
  - Phrase
  - Fraction of term
  - Symbol or acronym

- Kinds of Terms--anything on the data base
  - Subject terms appearing in titles, text, or as index terms
  - Author names

Kinds of Terms (cont'd)

Company names

Journal names as represented by the
standard ASTM CODEN

Country

CA section numbers

BA CROSS Codes

BA BIOSYSTEMATIC Codes

EI Card-A-Lert Codes

4.2.2 Truncation

Since many data bases include titles, which are author generated and therefore uncontrolled, it is necessary to include in one's profile all forms of a desired term to ensure retrieval of the desired information. In order to simplify this task of specifying all possible relevant word forms and fragments, CSC has allowed all options in truncation. Left, right, both, and none modes of truncation are permitted. When a search term is specified with no truncation, it requires an exact match with a term on the data base. Left truncation allows substitution of any prefix; right, of any suffix; and both, allows all of the preceding plus simultaneous substitutions of prefix and suffix on a term or term fraction. (See Figure 4-4). In addition to these four modes there is a fifth possibility, infix truncation, wherein substitution is allowed on an infix while prefix and/or suffix remain constant; we are considering the possibility of adding infix truncation to the CSC system. Figure 4-5 shows how it would be used.

Truncation can be used with any kind of data element or term type in a given data base. The usefulness of right truncation is usually readily understood. Right truncation is used to select singular, plural, and other forms of words that contain a common stem. In order to regularize the use of commonly

Mode	Function	Example
none	requires exact match of a term	term AZO
left	allows substitution of any prefix on the term	* term DI AZO
right	allows substitution of any suffix on the term	term * AZO XY
both	Allows substitution of any prefix and/or suffix	* term * DI AZO METHANE

NOTE: * denotes truncation

Figure 4-4
TRUNCATION MODES

truncated terms and to assist in the selection of optimal truncation forms, we have prepared a Truncation Guide for right truncated words. See Section 7 on User Aids for details. The use of right truncated terms is quite apparent. On the other hand, the usefulness of left truncation is not so obvious but it can be readily demonstrated. For example one might use the left truncated term *MYCIN to represent antibiotics and retrieve many relevant terms as can be seen in Figure 4-6.

The usefulness of the "both" truncation mode can be seen in the case where a user interested in organometallic compounds--especially those containing tin--might specify both left and right truncation by putting an asterisk on either side of the term tin in his profile. Thus, the search term *TIN* would retrieve the compounds: tetraphenyltin, triethyltin, and bistributyltinoxide.

When truncating, one has to be careful not to use term fragments or letter groupings that occur frequently in unrelated words. In order to avoid inappropriate truncations and identify beforehand those candidate search terms that might produce irrelevant hits, we have prepared a KLIC (Key-Letter-in-Context) Index* for each data base in use at IITRI. See Section 7 for details.

---

*Note: The KLIC Index was first developed at the University of Nottingham in England.

Infix truncation permits search for any variable fragment of a term between prefix and suffix.

A * B

Examples of its usefulness in chemical literature:

electron - * - resonance	would retrieve
electron - spin - resonance electron - paramagnetic - resonance	
tri * cobaltate (II)	would retrieve such compounds as
trioxalato cobaltate (II) trichlorocobaltate (II) triiodocobaltate (II)	
glucose - * - phosphate	would retrieve both
glucose - 1 - phosphate glucose - 6 - phosphate	

Figure 4-5

INFIX TRUNCATION

Use of the term *MYCIN for antibiotics retrieves

ACTOMYCIN
ANTIMYCIN
BIOMYCIN
ERYTHROMYCIN
NEOMYCIN
STAPHYLOMYCIN
STREPTOMYCIN

and many others

One search term *MYCIN substitutes for 20 to 30 specific terms.

Use of simultaneous left and right truncation would pick up all of the above terms plus the plural forms.

Figure 4-6

LEFT TRUNCATION

4.2.3 Linking or Grouping of Terms

In order to simplify the writing of a profile, similar or semantically related terms may be linked together in a single expression by a link code. Terms that are semantically associated can be linked together in a single expression. That is, several terms that are synonymous, related, or hierarchically broader and narrower, can be represented by a single alphabetic character. This simplifies the user's task of writing his logic expression. He can merely specify a link designator rather than indicate the multiple terms joined by the link in cases where any one of the terms would be equally satisfactory in the logic expression. For example, a user interested in reactions of halogens and alkali metals would use the terms listed below and assign the link codes "A" and "B".

Terms	Link Code		Terms	Link Code
Halogen	A		Alkali metals	B
Halide	A		Lithium	B
Fluorine	A	AND	Sodium	B
Chlorine	A		Cesium	B
Bromine	A		Potassium	B
Iodine	A		Rubidium	B

In writing his logic expression he would not have to specify the terms:

(Halogen | Halide | Fluorine | Chlorine | Bromine | Iodine)

and

(Alkali metals | lithium | sodium | cesium | potassium | rubidium)

He can merely specify

(A & B)

4.2.4 Logic

An effective profile requires not only the use of appropriate search terms but also that the terms be related to each other in a manner that correctly represents the intent of the search question. The relationships are expressed in the algebra of logic, called Boolean algebra. Three logic operators are used to indicate the relationships between search terms: AND, OR, and NOT. The logic symbols used are as follows:

Logic Operators	Symbol
AND	&
OR	|
NOT	¬

AND logic, designated &, will retrieve an item only if both terms connected by the AND operator are present.

The & operator is the familiar conjunction or intersection of mathematics and engineering in which it can be represented by x, ', $\cap$, or $\wedge$.

OR logic, designated |, will retrieve an item if either one or both the terms connected by the OR operator are present.

The | operator is the familiar inclusive disjunction or union of mathematics and engineering in which it can be represented as +, $\cup$, or $\vee$.

NOT logic, designated ¬, will cause items containing a term designated by the NOT operator to be rejected.

The ¬ operator is also referred to as complement or negation and can be represented by $\overline{}$ (overline) or ' .

Because NOT is a unary operator relating to only one term, it is necessary to always precede the NOT operator with an AND operator in writing a logic expression. Thus the logic expression for a search for a compound having no nitrogen and containing oxygen or carbon would be written as:

oxygen OR carbon AND NOT nitrogen.

Parentheses can be used to limit the effect of the NOT term. In the expression

$$A \,\&\, (B \mid C) \,\&\, \neg D$$

if D is present, the entire expression is false. In the modified expression

$$A \,\&\, (B \mid (C \,\&\, \neg D))$$

the expression will be true if A and B are present even if D is also present.

Terms connected by AND or OR are not affected by sequence. Thus,

A & B = B & A
A | B = B | A.

Similarly, AND or OR are not affected by grouping. Thus,

(A & B) & C = A & (B & C)
(A | B) | C = A | (B | C).

It should be noted, however, that the placement of parenthesis in a mixed expression can alter the logic. A & (B | C) is not the same as (A & B) | C.

Several laws of logic may be helpful in determining the consequences of writing elementary logic expressions.

By the law of absorption:

A & (A | B) = A
A | (A & B) = A.

By the law of distribution:

A & (B | C) = (A & B) | (A & C)
A | (B & C) = (A | B) & (A | C).

By the law of duality:

¬(A & B) = ¬A | ¬B
¬(A | B) = ¬A & ¬B.

The logic can be written in any free form Boolean expression. To avoid logical ambiguity, however, parentheses should be used freely. There is no restriction on the number of parentheses used; care should be taken to ensure that the number of left parentheses equals the number of right parentheses. The logic expressions for profiles can be as specific and involved as is necessary to express the user's question. While most expressions are relatively simple, any expression can be handled by the system. For example, the following expression would be legitimate:

$$(((A\&B) \mid (C \mid D \mid E \mid F)) \& \neg G) \mid ((H\&I) \& \neg J)$$

However, experience indicates that useful retrieval can be achieved with a simple logic expression, whereas an overly complex expression may obscure a question and result in poor retrieval.

4.2.5 Weights

CSC profiles permit the assignment of weights by users to further refine their profiles. Weights are numerical values from 0 to 9 assigned to terms to specify their relative importance. If a user employs weights in his profile the output is arranged in descending weighted order so that those citations with the highest weights--presumably the references that are of most significance to the user--will be on top. Since the output of a search will be limited to the specified maximum number of hits, the printed output can include the highest ranked weights above the cut-off number of hits. If the user chooses not to use weights (and this is usually the case), the output is ordered either numerically by citation number or alphabetically by author (first letter of the first author's last name).

Although the designed purpose of weights was to allow further specification in a given profile, CSC has found that users employ weights in order to separate two or three profiles that are submitted as one profile for one subscription fee.

4.3 Profile Format

After a profile has been written and checked by the profile coordinator it is keypunched. The keypunched profile consists of a header card, a group of term cards, and one or more cards containing a logic expression. These cards have the following internal structure:

Card	Columns	Contents
Header	1-10	Profile number
	11-13	Number of terms
	14-16	Number of links (a link is a group of disjoint terms)
	17-19	Maximum number of cards to be printed
	20-22	Minimum number of terms necessary to satisfy the logic expression
	23-25	Private Libraries usage (contains 'PRI' if output is to be placed in a Private Library)
	26	Output medium (C=cards P=paper)
	27	Number of copies (b̸=1, else 1-9 permissible)
	28-29	Sort type for output (AN=ascending citation number order, WT=descending weight order, 03=author)
Term	1-10	Profile number
	11-13	Term number
	14	Truncation mode (0=none, 1=left, 2=right, 3=both)
	15-16	Type of field to be tested
	17	Link (terms with the same letter in this place are OR'd together. The link letter may then be used as an operand in the logic expression)
	18	Weight of term (if weights are used)
	19-38	The term
	49	On last term card this position is '1'
Logic	1-10	Profile number
	11-12	Minimum number of terms that must be found to satisfy the logic

Card	Columns	Contents
Logic	13-59,60	The logic expression, consisting of terms (3-digit term numbers), links (single characters), Boolean operators ("&", '|', '¬'), and parentheses. If the expression takes more than one card, all but the last card have a '1' in position 60. The last character of the logic expression is '$'.

4.4 Profile Modification

The problem of preparation and modification of search profiles has undergone careful investigation at the Center in light of the relevant statistics and the summary of experience obtained from the pilot group of users. The best profile is prepared when the person writing the profile understands three things: the intent and terminology of the search problem, the contents of the data file, and the characteristics of the search system. Ideally, the user who has the best understanding of his problem should become familiar with the contents of the data file and the search system so that he can write an effective profile. In lieu of that, if he is unwilling or unable to do so, the responsibility is assumed by a middleman either at the user's institution or at the Center.

At CSC we have handled profiles prepared in all three ways. As would be expected, in the cases where the user took sufficient interest to learn the system and write and modify his own profile, the result was a good profile and a satisfied user. Good results were also obtained when the user took sufficient time and interest to fully explain his search problem to a company or Center profile coordinator.

In preparing a profile for an SDI run, care must be taken to include not only the terms that describe the user's interests but also all synonyms for those terms used in the vocabulary of the particular data base to be searched. Omission of similar terms may result in a loss of pertinent articles. A logic expression combining those terms must also be developed that will not be too general or too restrictive. Since the preparation of a profile for an initial run may not completely describe the user's interest, it is usually necessary to modify the profile three times to correct omissions of terms and faulty logic.

The output produced for the first few runs of a new profile can be reviewed to help modify the profile. Several questions must be considered in making revisions in the profile. Are all pertinent articles retrieved by the SDI run? This can be answered by a comparison with a manual search of the material covered by the SDI run. If there are missing articles, the omitted citations must be studied for additional terms and logic to be added to the profile. The terms may be present in the profile, but the logic may be restrictive. In this case, the logic must be relaxed, but at the same time, not overly generalized. Is the SDI run producing a great deal of nonpertinent material? This may be due to inclusion of terms that are too general, for example, ENZYME may be used when the names of specific enzymes would bring about more relevant retrieval. The logic expression may also be too general and need to be more restrictive. These terms might fall into the classification of positive hit terms tied to the logic by the AND logic operator or modifying terms may need to be of the negative type. These terms would cause a citation to be rejected if the negative words appeared in the citation.

It is possible that some questions submitted to an SDI system are of such a nature that much nonpertinent material must be retrieved in order to gather the citations that are of definite interest. In contrast, it is also possible that a subject may be so new or esoteric that little has been published. This type of question may legitimately produce very small quantities of output with very few articles of real interest.

Based on our experience between September 1969 and June 1972 and our observation of user preparation and modification of profiles, we have come to the conclusion that although users can be trained to write their own profiles, the user who conscientiously revises and updates his own profile under his own impetus is the exception rather than the rule. CSC experience indicates that since it requires almost as much time to check a user-written profile as to write it, it would be more advantageous to write the original profiles. CSC would then be in a better position to revise profiles for the users. CSC profile coordinators are closer to the data bases, can recognize data base content changes more rapidly than the users can, and hence can respond by changing profiles accordingly.

Several user aids have been prepared by the Computer Search Center to assist the staff and users in developing, evaluating, and modifying search profiles. These are described in Section 7--User Aids.

5. SOFTWARE SYSTEM

The CSC software system was designed to accomodate a variety of types of users with a variety of types of data bases that would meet their needs. Search programs for handling machine-readable data bases are expensive to develop and expensive to maintain. Since we had no desire to incur the expense of maintaining multiple search programs, we developed a general purpose search program that would handle virtually any of the machine-readable data bases containing natural language information.

When handling multiple data bases, one is very likely to encounter multiple character sets and multiple character codes. The tape formats and record formats differ from data base to data base. In fact, they differ within data bases that are produced by the same organization. The data elements contained on the tapes vary considerably from tape to tape. This format variation problem that occurs when handling multiple data bases led to the adoption of the standard IITRI file structure and preprocessor system described in Section 5.1.

The general purpose search system carries out the five basic functions of preparing profile input, preparing data base input, searching the data base for information corresponding to the user profiles, preparing output for dissemination to the users, and maintaining statistics. These are shown in a generalized flow chart, Figure 5-1.

The five basic programs consist of eleven program modules. Descriptions of the main programs, constituent program modules, and the files by which they communicate with each other are presented in Sections 5.3 and 5.5. Flow charts showing the interrelationships and interfaces between and among programs and files are presented in Figures 5-2, 5-3, 5-4, 5-5, 5-6, and 5.7. The development of the CSC search strategy is discussed in Section 5.6 and logic systems--current and projected--are presented in Section 5.7.

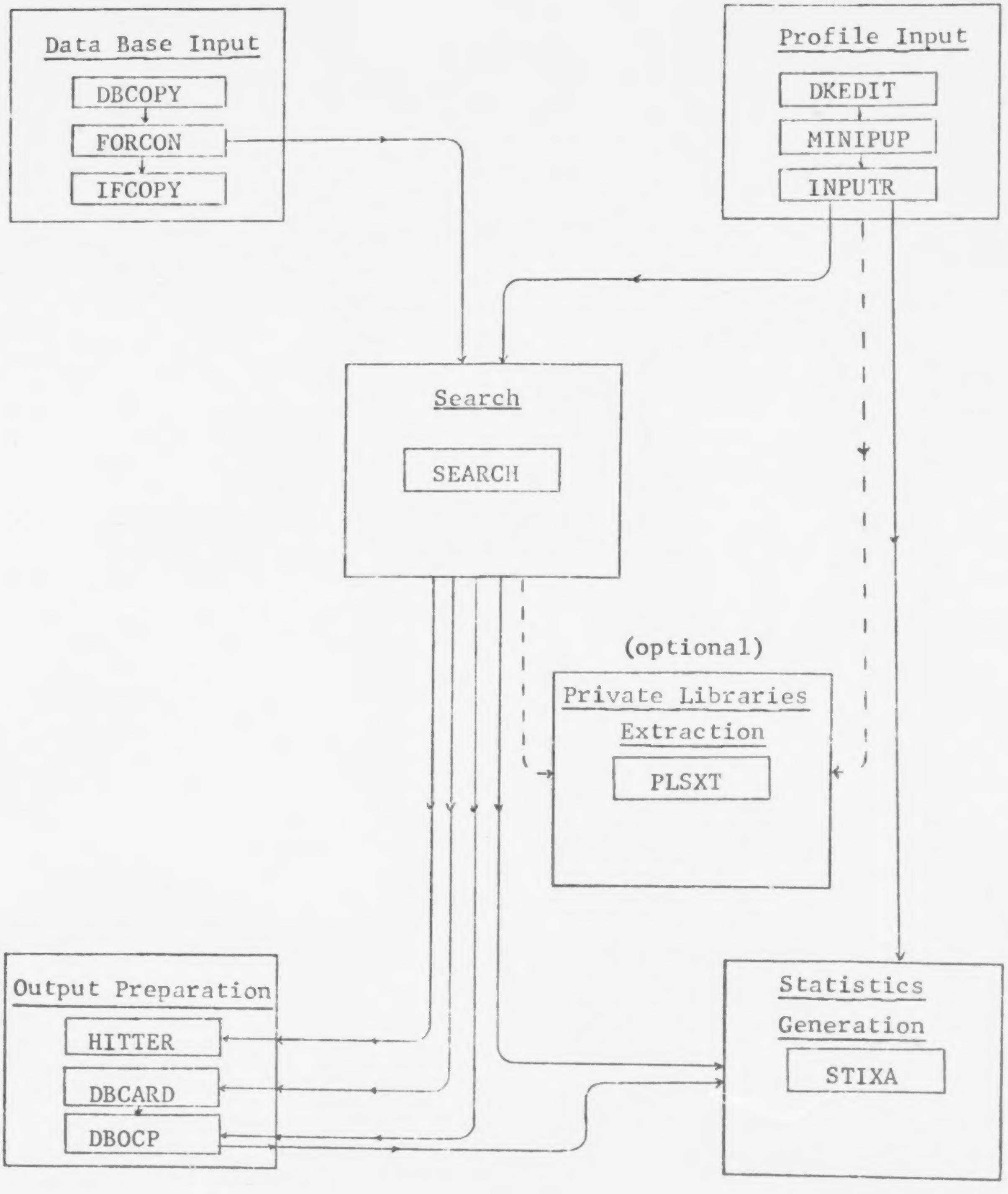

Data Base Input
DBCOPY
FORCON
IFCOPY
Profile Input
DKEDIT
MINIPUP
INPUTR
Search
SEARCH
(optional)
Private Libraries
Extraction
PLSXT
Output Preparation
HITTER
DBCARD
DBOCP
Statistics
Generation
STIXA

Figure 5-1
SDI SYSTEM GENERALIZED FLOW CHART

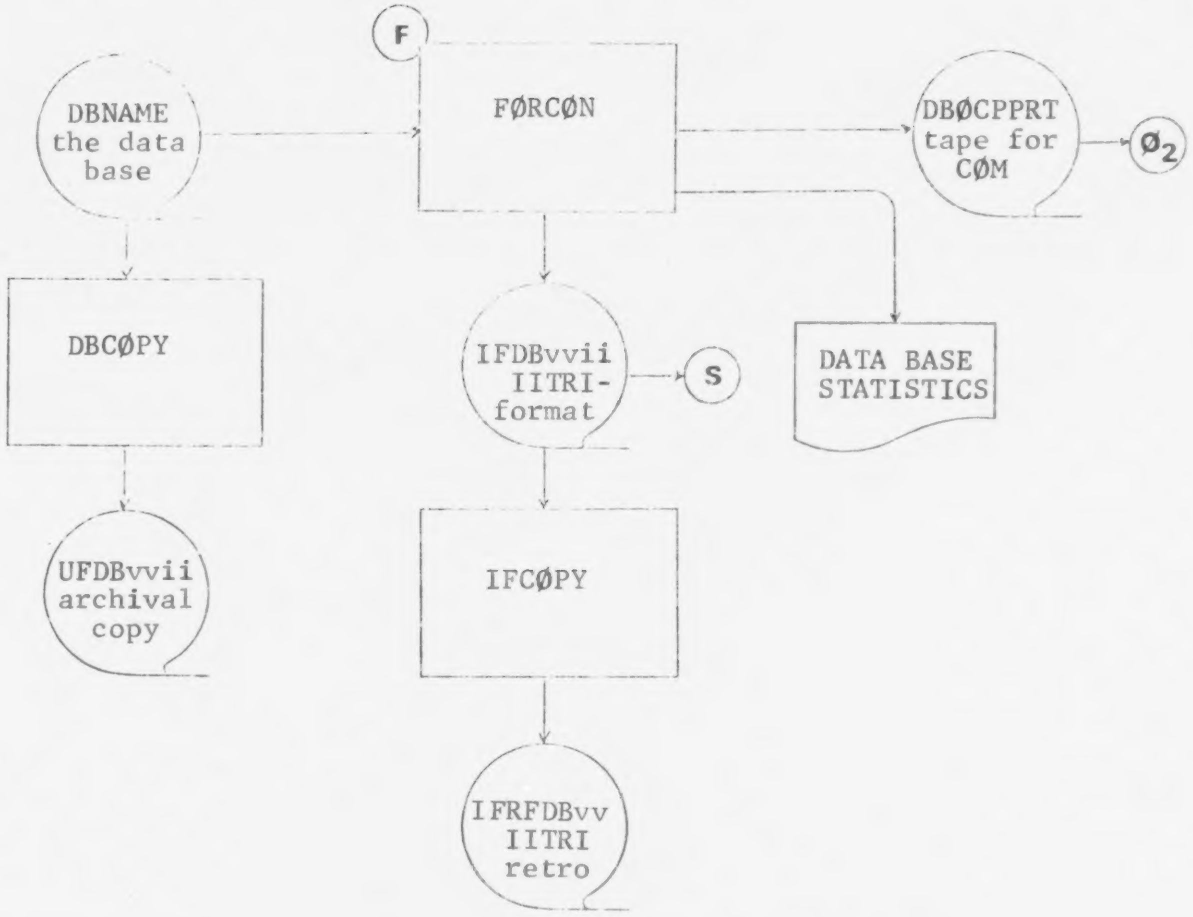

F
DBNAME
the data
base
FØRCØN
DBØCPPRT
tape for
CØM
Ø2
DBCØPY
IFDBvvii
IITRI-
format
S
DATA BASE
STATISTICS
UFDBvvii
archival
copy
IFCØPY
IFRFDBvv
IITRI
retro

Figure 5-2
SDI SYSTEM DETAILED FLOW CHART
PART 1: DATA BASE INPUT

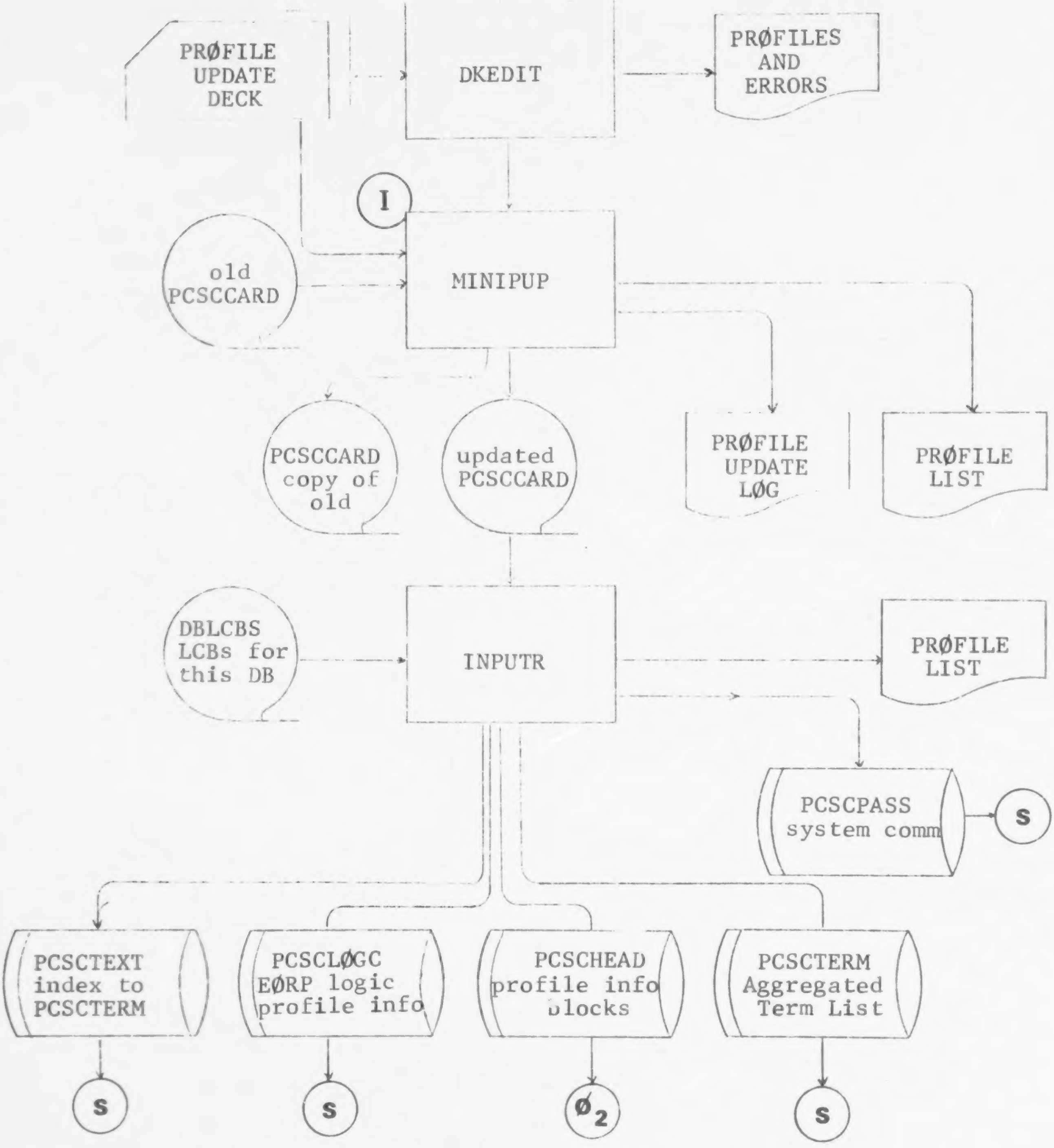

PROFILE UPDATE DECK
DKEDIT
PROFILES AND ERRORS
I
old PCSCCARD
MINIPUP
PCSCCARD copy of old
updated PCSCCARD
PROFILE UPDATE LOG
PROFILE LIST
DBLCBS LCBs for this DB
INPUTR
PROFILE LIST
PCSCPASS system comm
S
PCSCTEXT index to PCSCTERM
PCSCLOGC EORP logic profile info
PCSCHEAD profile info blocks
PCSCTERM Aggregated Term List
S
S
0_2
S

Figure 5-3

SDI SYSTEM DETAILED FLOW CHART

PART 2: PROFILE INPUT

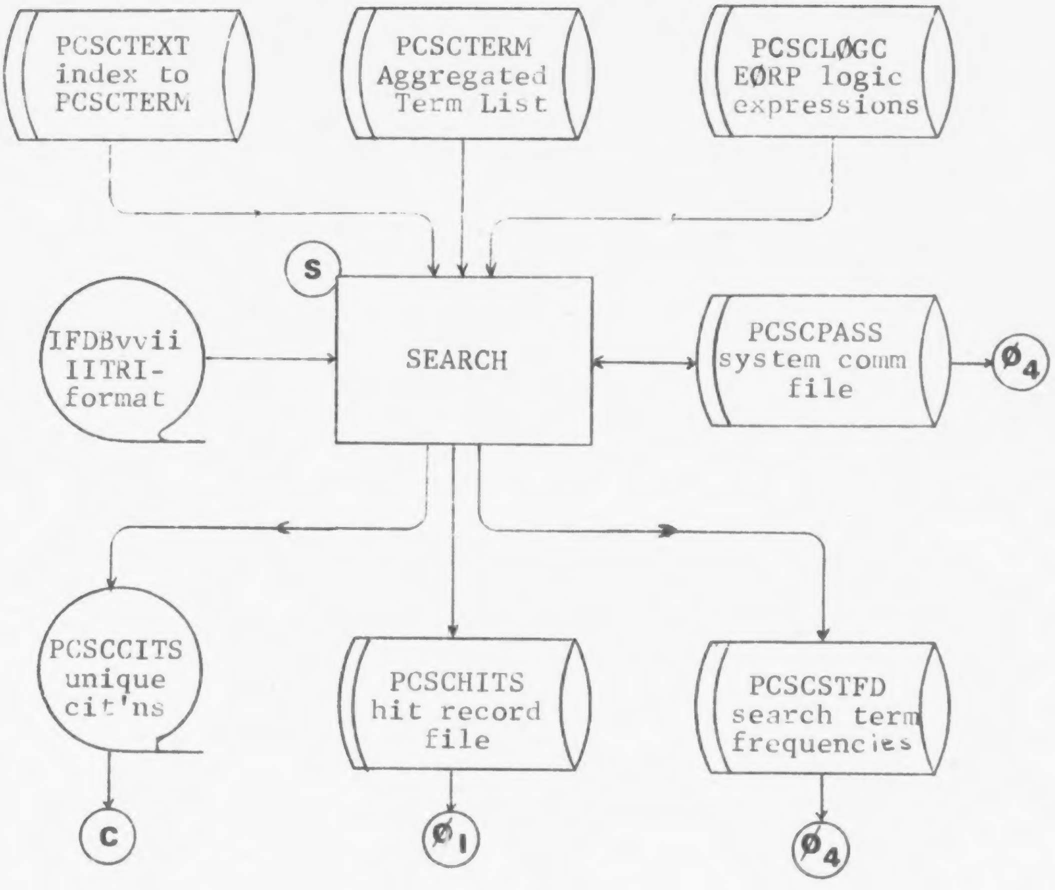

PCSCTEXT
index to
PCSCTERM
PCSCTERM
Aggregated
Term List
PCSCLØGC
EØRP logic
expressions
S
IFDBvvii
IITRI-
format
SEARCH
PCSCPASS
system comm
file
Ø4
PCSCCITS
unique
cit'ns
PCSCHITS
hit record
file
PCSCSTFD
search term
frequencies
C
Ø1
Ø4

Figure 5-4
SDI SYSTEM DETAILED FLOW CHART
PART 3: SEARCH

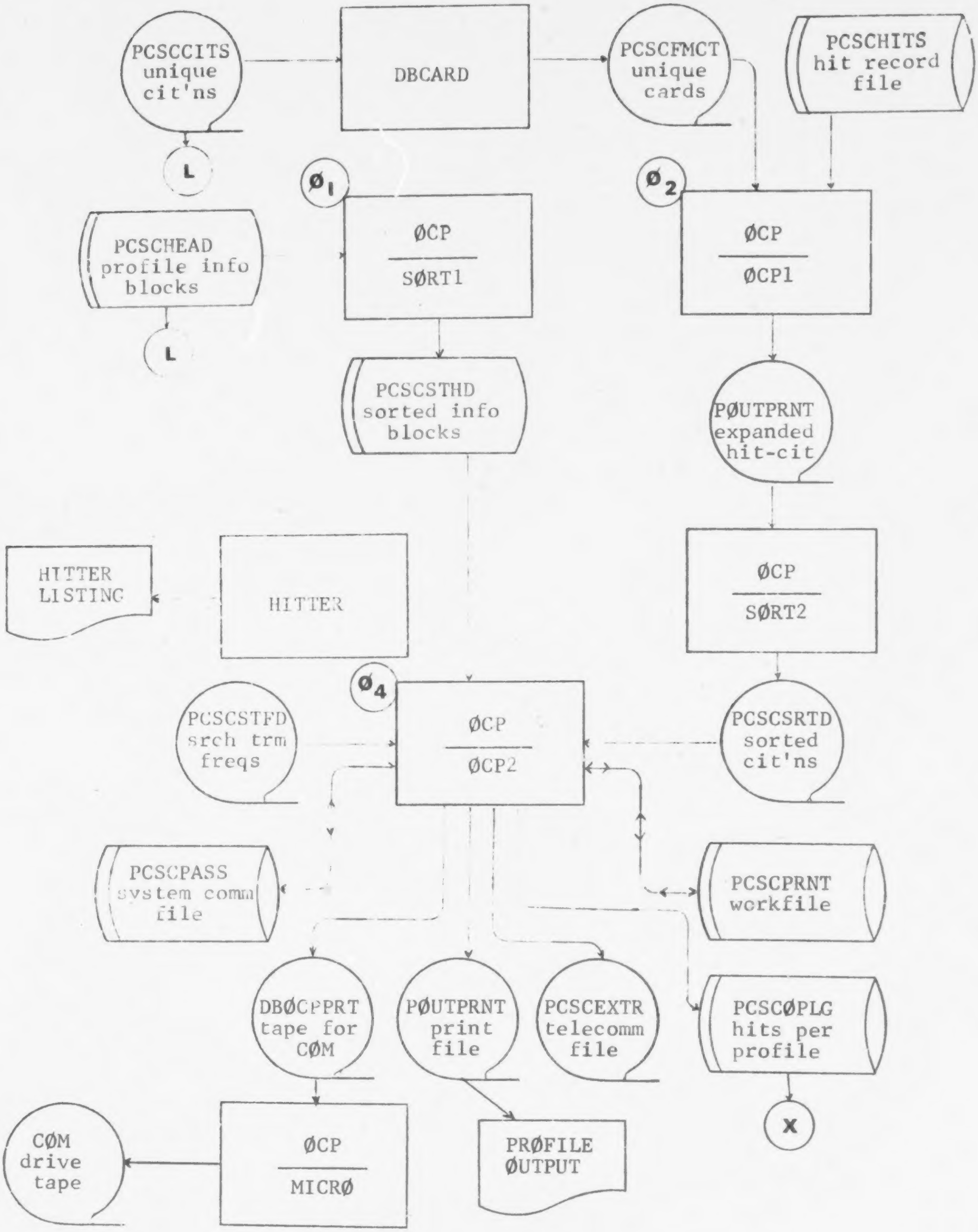

PCSCCITS unique cit'ns
DBCARD
PCSCFMCT unique cards
PCSCHITS hit record file
L
Ø1
Ø2
PCSCHEAD profile info blocks
ØCP
SØRT1
ØCP
ØCP1
L
PCSCSTHD sorted info blocks
PØUTPRNT expanded hit-cit
HITTER LISTING
HITTER
ØCP
SØRT2
Ø4
PCSCSTFD srch trm freqs
ØCP
ØCP2
PCSCSRTD sorted cit'ns
PCSCPASS system comm file
PCSCPRNT workfile
DBØCPPRT tape for CØM
PØUTPRNT print file
PCSCEXTR telecomm file
PCSCØPLG hits per profile
X
CØM drive tape
ØCP
MICRØ
PRØFILE ØUTPUT

Figure 5-5

SDI SYSTEM DETAILED FLOW CHART

PART 4: OUTPUT

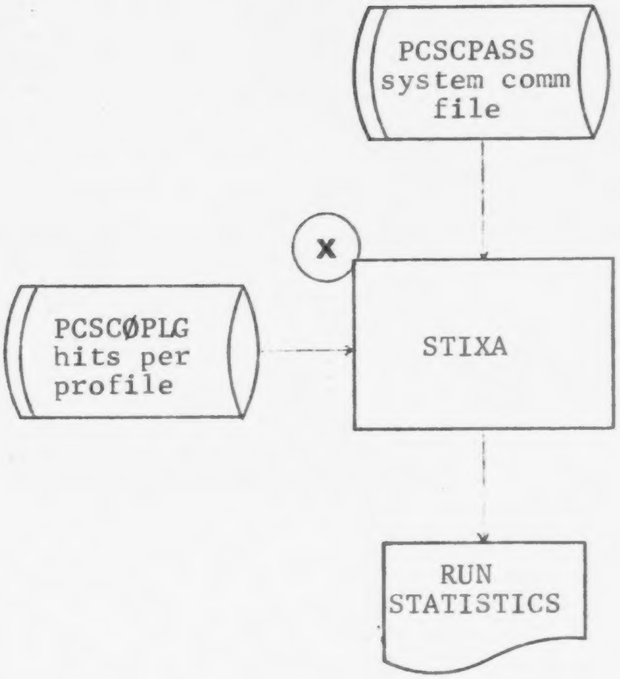

PCSCPASS
system comm
file
x
PCSCØPLG
hits per
profile
STIXA
RUN
STATISTICS

Figure 5-6
SDI SYSTEM DETAILED FLOW CHART
PART 5: STATISTICS

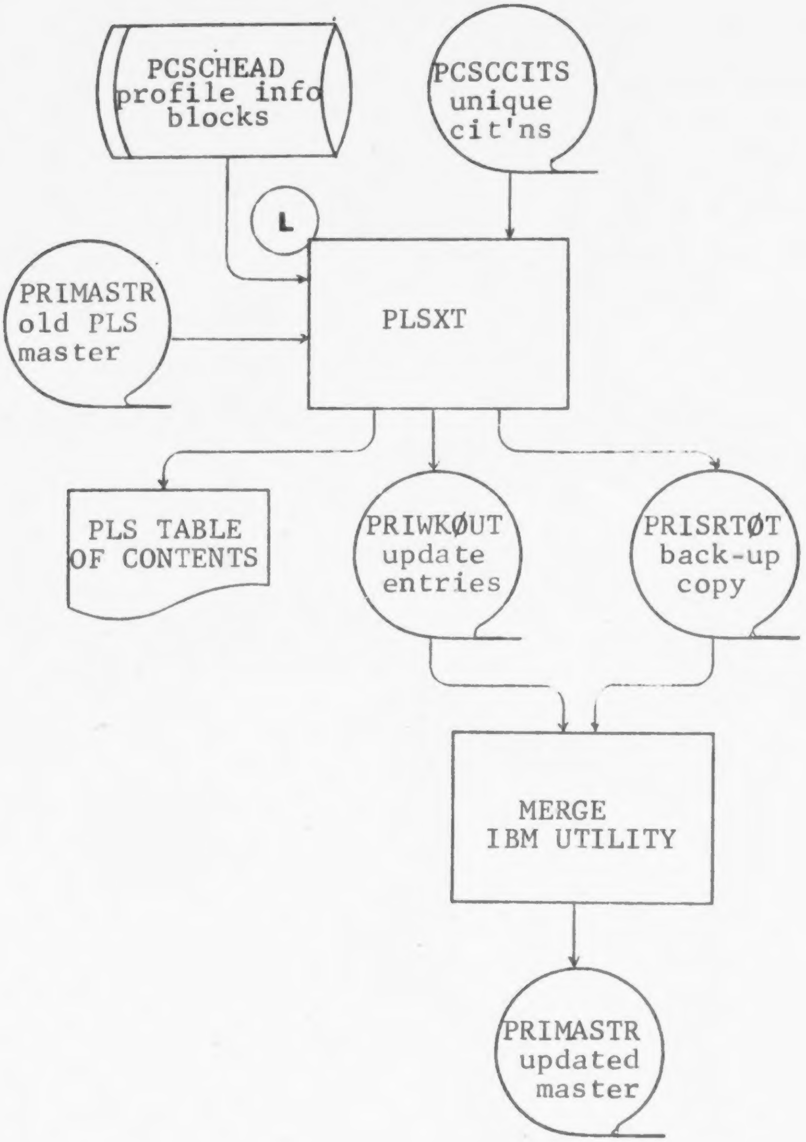

PCSCHEAD
profile info
blocks
PCSCCITS
unique
cit'ns
L
PRIMASTR
old PLS
master
PLSXT
PLS TABLE
OF CONTENTS
PRIWKØUT
update
entries
PRISRTØT
back-up
copy
MERGE
IBM UTILITY
PRIMASTR
updated
master

Figure 5-7
SDI SYSTEM DETAILED DATA FLOW
PART 6(Optional): PRIVATE LIBRARIES SYSTEM EXTRACTION

A Private Libraries System (PLS) that is not a basic part of the CSC software system is discussed in Section 5.8. It is a generalized system for creation and maintenance of user-defined private files that may contain virtually any document records the user wants to retain. PLS interfaces with the CSC system and can automatically accept as input specified output from the CSC system. This is another example of the benefits of modular programming.

5.1 Data Base File Structure and Preprocessor System

The requirement for a single generalized programming system for processing multiple data bases necessitated the design of a file structure that would accommodate all of the variables one might encounter in different data bases, such as multiple character sets and character codes, differing tape formats and record formats, different data elements and different ways of representing the same data element. In the IITRI system a different data type code is assigned to each kind of data element found on a data base. The data elements found in the data bases we are now using are shown in Figure 5-8.

Each data base that is to be searched is reformatted by a preprocessor program that converts the tape into our standard file structure. (See Figure 5-9.) After reformatting, each record is composed of a key, directory, and character string. The key contains the volume, issue, and citation number as given by the data base supplier, and the directory identifies each type of element contained in the record according to IITRI data type codes. The string contains the data.

In the directory the data type code is followed by the starting position for the actual data and an indication of the number of characters required by the data. Thus, in Figure 5-10 for the record having Citation Number 81368 of Volume 74, Issue 16, in Chemical Abstracts Condensates there is a CODEN that starts in position 1 and is 26 characters long. The next kind of data element included in the record is a Journal Name which has a data type code "04". The actual data starts in position 27, one position beyond the end of the CODEN data, and is 14 characters long. The next data element is the title which has data type code "02" and starts in position 41, one position beyond the end of the journal data. The title data is 76 characters long, and the rest of the data are recorded in a similar fashion. Following on through Figure 5-10, the format becomes obvious. The string portion of Figure 5-10 shows how the actual data for this particular reference is contained in IITRI format on tape and the complete record, which appears in the lower portion of Figure 5-10, shows the entire key, directory and character string for the particular record as it appears on tape.

The use of data element codes allows us to handle multiple, varied data elements. The system also allows us to add new data elements and new data type codes as they arise. We have no way of knowing what new data elements suppliers may include in their tapes a few years from now. However, we have allowed for 2^{35}-1 different data type codes. It is unlikely that we will be unable to accommodate any new data element that may come into existence.

The standard IITRI format is employed for any data base processed. Our method for handling multiple data bases is to write a preprocessor program for each different data base that is handled in the system. The preprocessor program reformats the data that is contained on the supplier data base and puts it into IITRI format. In that way every data base looks the same to the search program, and all data bases can be handled by one and the same search program.

The preprocessor or format conversion programs are referred to in the CSC system as FORCON programs. Details regarding the development of the format conversion programs for a variety of data bases are given in the following section.

Data Element	IITRI Data Type Codes
Source information	01
CODEN	
Journal reference	
Pagination	
Dates	
Title of article	02
Author(s)	03
Short journal title	04
Keyword(s)	05
Index terms	
CA section number	
CA Registry number	06
Molecular formula	07
Corporate author	08
Abstract text	09
BA CROSS code	10
BA biosystematic index	11
EI Card-A-Lert Code	12
Publication information	13
Original language	
Availability	
Publisher	
Price	
Parent journal	
Original abstract source	
CA cross reference	14
Patent priority class	15
Secondary source	16

Figure 5-8
DATA ELEMENTS AND IITRI DATA TYPE CODES

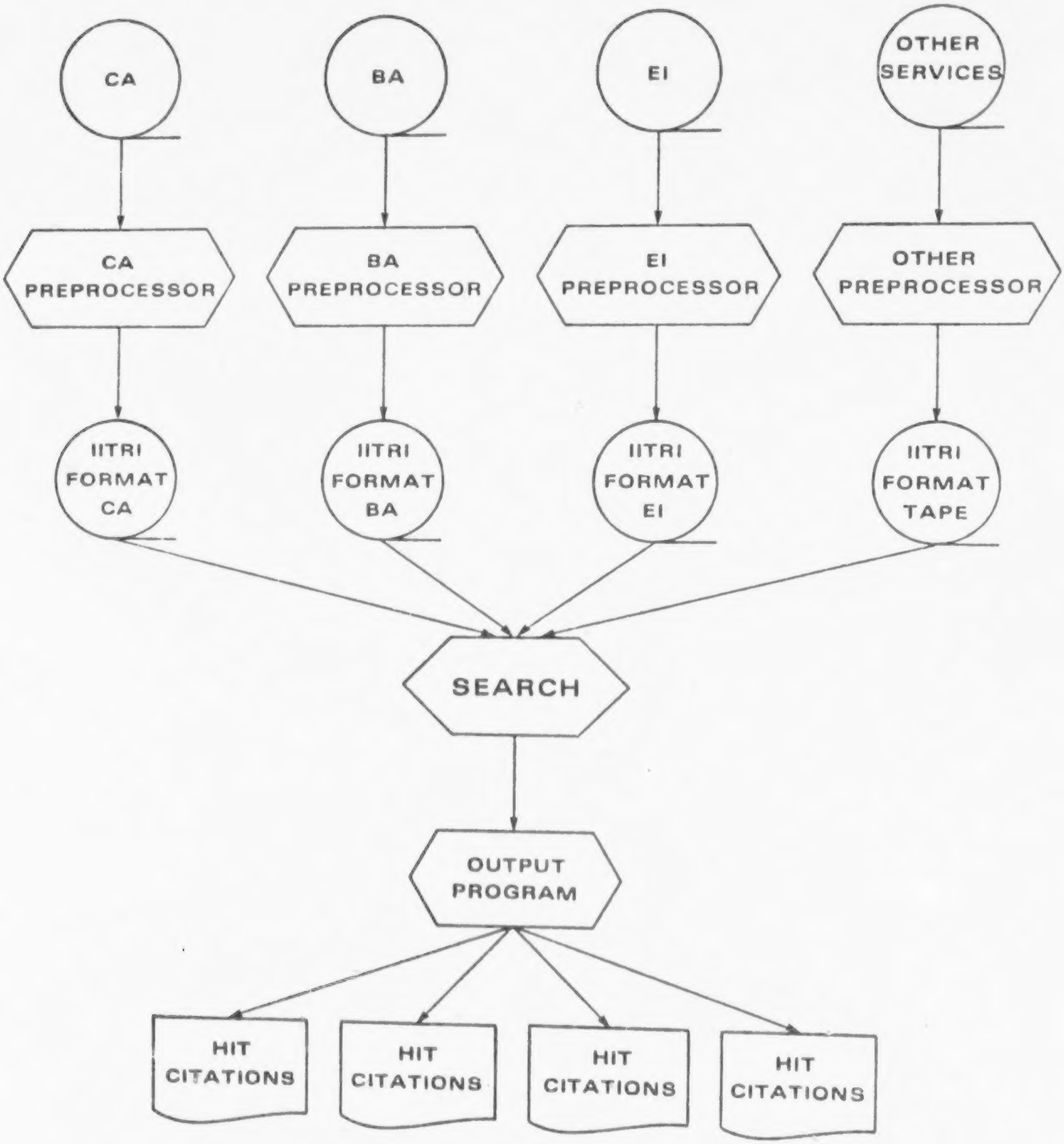

CA
BA
EI
OTHER SERVICES
CA PREPROCESSOR
BA PREPROCESSOR
EI PREPROCESSOR
OTHER PREPROCESSOR
IITRI FORMAT CA
IITRI FORMAT BA
IITRI FORMAT EI
IITRI FORMAT TAPE
SEARCH
OUTPUT PROGRAM
HIT CITATIONS
HIT CITATIONS
HIT CITATIONS
HIT CITATIONS

Figure 5-9
PREPROCESSOR SYSTEM

Key: 7416-081368 (Volume, Issue and Abstract Number)

Directory:

1	1	26	(CODEN)
4	27	14	(Journal)
2	41	76	(Title)
3	117	60	(Author(s))
8	177	51	(Corp. Author)
5	228	40	(Index Terms)
13	268	17	(Language)

String:

```
JPCHAX/75/3/325-30/000071/J.  PHYS. CHEM.VIBRONIC EFFECTS IN
THE INFRARED SPECTRUM OF THE ANION OF TETRACYANOETHYLENEDEVLIN,
J. PAUL$MOORE, JESSE C.$SMITH, DONALD$YOUHNE, YOUNG$DEP. CHEM.,
OKLAHOMA STATE UNIV., STILLWATER, OKLA.$CA073000$ IR SPECTRA
ALKALI METAL SALTSORIG. LANG.:  ENG
```

Complete Record Appears on Tape as:

```
7416-081368    1    1    26    4    27    14    2    41    76
3   117    60    8    177    51    5    228    40    13    268
17 JPCHAX/75/3/325-30/000071/J. PHYS. CHEM. VIBRONIC EFFECTS
IN THE INFRARED SPECTRUM OF THE ANION OF TETRACYANOETHYLENEDEVLIN,
J. PAUL$MOORE, JESSE C.$SMITH, DONALD$YOUHNE, YOUNG$DEP. CHEM.,
OKLAHOMA STATE UNIV., STILLWATER, OKLA.$CA073000$ IR SPECTRA
ALKALI METAL SALTSORIG. LANG.: ENG
```

Figure 5-10
IITRI FORMATTED CITATION

5.2 Format Conversion Programs

5.2.1 Variability of Data Base Format

In the course of developing the CSC system we have examined numerous data bases, both to determine the feasibility and cost of converting them to our format for searching and to determine whether sufficient user interest exists to warrant marketing them. Among those we have studied are:

- Biological Abstracts Previews (BAP)
- Chemical Abstracts Service (CAS) data bases:
  - Condensates
  - Integrated Subject File (ISF)
  - Chemical Industry Notes (CIN)
  - Chemical Titles (CT)
  - Chemical-Biological Activities (CBAC)
  - Polymer Science and Technology (POST)
- Engineering Index (EI) COMPENDEX
- Educational Resources Information Center (ERIC)
- Food Science and Technology Abstracts (FSTA)
- Government Reports Announcements (GRA)
- Institute for Scientific Information (ISI)
- Institution of Electrical Engineers (England) (INSPEC)
- Medical Literature Analysis and Retrieval System (MEDLARS)
- Metals Abstracts Index (METADEX)
- Searchable Physics Information Notices (SPIN)

Further information on these and other machine-readable data bases is contained in the Association of Scientific Information Dissemination Centers (ASIDIC) Survey of Information Center Services.[3]

Despite several proposed standards for tape data bases, including those of the Committee on Scientific and Technical Information (COSATI), the American National Standards Institute (ANSI), and the International Standards Organization (ISO), no one format is in general use. Several of the publically available data bases are "based on" standards, but none can claim exact adherence. Many data bases have adopted the directory-plus-string organization, but organization and contents of the directory, data tag values, character codes, and control information vary widely. Since most standards do not include data element tag values (the codes which specify the contents of a given field), even those data bases designed around the same standard may use widely different codes. Some suppliers have designed hierarchies of codes (e.g., in the INSPEC data base, the type 3xx data elements are identification codes such as 310 for CODEN, 320 for ISBN, etc.) while others assign codes in random fashion (e.g., CAS uses sequentially assigned numbers to handle new data types). Since the standards include a header that describes the format of the directory, not only the code values but the code formats can differ. One supplier might use a three-digit numeric code and another a five-digit code. Some suppliers, however, have not adopted the directory plus string organization. The ISI data base involves fixed-format records, with the attendant complications necessary to allow varying length data. The CAS data bases, which share a format among themselves, use a modified directory plus string organization, but also allow short items to be stored in the directory itself. In addition, even data bases within the CAS Standard Distribution Format (SDF) have significant variations. Most CAS data bases use the same data element, the Temporary Abstract Number, to associate the physical records describing a single citation into a single logical record. The CAS-CIN data base, however, does not give Temporary Abstract Numbers at all, but uses a different data element to make the necessary association.

5.2.2 Data Base Documentation

In view of the wide diversity in data base formats it is particularly unfortunate that documentation is not very good. Although some suppliers, such as CAS and INSPEC, provide complete and detailed information along with examples and printouts, other sample tapes have been received with documentation as crude as a six-page Xeroxed description. Often, too, it is the data base which is poorly designed or overly complex that comes with the least satisfactory documentation.

5.2.3 Programming

In order to search a given data base, we first write a program to convert the supplier's tapes into our format. IITRI format is a directory-plus-character-string organization, using pure binary values in the directory and an EBCDIC-coded character string. While this mixed-mode arrangement is undesirable for distribution of a data base, it allows much faster access to data during processing. Since our format is used only for our internal purposes and not for distribution, we can justify this somewhat inelegant usage. If we were to distribute search output to users or other centers in magnetic tape form (currently prohibited by supplier license restrictions), we would convert all binary numbers to EBCDIC prior to distribution. CAS uses a similar mixing of binary tags and ASCII data on their distribution tapes and this mixture of storage modes makes hardware translation of ASCII to EBCDIC impossible. We then must expend a significant part of our conversion time for that data base on software translation.

The conversion from supplier format to IITRI format is done by a separate program for each data base. So far no two data bases have been found to be exactly compatible. Generally, however, the process of adding a new data base to our capability is simple. Most directory-plus-string data bases are similar enough that a new format conversion (FORCON) program can be based on an existing one. The changes necessary to convert a FORCON for USGRDR into one for INSPEC, for example, are relatively minor, since they are based on very similar standards. Data element tags and storage formats change, but the basic processing flow is unaltered. Also data bases from a single supplier may be very similar. The various CAS data bases in SDF can be handled by very modest changes in the conversion program. In the case of

CAS, the SDF data types are the same for all the data bases, except for a few types unique to single data bases, and storage formats are identical.

The task of writing a format conversion program has two parts. The first, understanding the data base, is always the more difficult. The actual writing of the program is almost trivial once we are thoroughly familiar with the data base. There are four stages of development in acquiring a new data base capability:

Stage 1	Evaluate the contents and format to determine complexity of conversion and usefulness of data.
Stage 2	Implement a rough conversion program to allow test search and production of samples.
Stage 3	Improve the Stage 2 FORCON for detailed testings to allow rough timing estimates and extended-period tests.
Stage 4	Implement a production FORCON, smoothing out logic and aiming at improved execution speed.

In many cases the results of Stage 1 or Stage 2 indicate that no further development is desirable at present. At this point we have the knowledge necessary to produce a FORCON or do basic tests if user interest develops, but no further work would be profitable either because user interest is negligible or implementation problems are unworkably large.

If a data base seems to have potential for CSC and Stage 1 and Stage 2 experience indicates a good data base and a satisfactory supplier, then a Stage 3 FORCON is a good investment and an extended trial is carried out. The EI COMPENDEX tapes, for instance, were tested for a year before we made a firm commitment to maintain subscriptions. Sometimes Stage 2 can be skipped. New CAS data bases, for instance, can be handled with such minor changes to existing FORCONs that virtually no preliminary testing need be done. The evaluation stage is also drastically reduced in such cases. A production FORCON is based on significant experience with the data base and incorporates changes and improvements designed to improve operating speed and consistency of output. At this point variations from the documentation, which virtually always exist, can be corrected. Also at this point, special output programs and card formats can be fixed, while earlier tests are done with standard or slightly-modified ones.

5.2.4 Status

Currently we have production-level FORCONs for CA Condensates (SDF), BA Previews, BioResearch Index, and EI COMPENDEX. These are well-tested programs and their logic flow and object code have been carefully analyzed for efficient operation. Test level FORCONs have been written for CBAC (pre-SDF), POST (pre-SDF), CT (pre-SDF and SDF), CIN (SDF), ISI, FSTA, and INSPEC. These programs have been tested and output has been checked for consistency and correctness. We are currently evaluating CIN for CAS. We plan to offer FSTA beginning in the fall of 1972. INSPEC and ISI are being evaluated for marketability. Evaluation-level FORCONs have been written for USGRDR, American Mathematical Society, ERIC, and many other data bases. These are being reviewed for suitability of contents and difficulty of conversion. Completeness of data is also checked (lack of CODEN, corporate author, or other data is a drawback).

The development of FORCONs and evaluation of data bases is a continuing part of CSC's development program. The resulting awareness of features of various data bases is useful in evaluation of our own system and in counseling our subscribers as well as in planning for future expansion to other data bases. In addition we can suggest desirable features from data bases we evaluate to suppliers of our production data bases. In some cases the suppliers are able to add features or revise procedures on the basis of our suggestions.

5.3 Program Descriptions

The SDI system consists basically of a group of programs for handling data bases. The programs communicate with each other by means of data files (see Section 5.5). There are five basic programs for carrying out the five basic functions for data base input preparation, profile input preparation, search, output preparation, and statistics generation. These programs consist of a number of modular programs and subroutines. The constituent programs are described below together with an indication of the files they use.

5.3.1 DBCOPY

DBCOPY (Data Base COPY program) copies the data base tape in the supplier's format to another tape for archival storage. Six CA tapes are stored on each archive tape.

File Name	Use in DPCOPY
DBNAME	The data base tape. Scratched after copying.
UFDBvvii	The archival copy (vv = vol, ii = issue). Kept permanently.

5.3.2 FORCON

FORCON (FORmat CONversion program) reads the data base tape and converts it from the supplier's format to IITRI format. All records dealing with a citation are read in turn and the IITRI-format directory and a preliminary string are assembled. The final string is formed by extracting portions of the preliminary string, performing any necessary translations, and concatenating them. The translation and concatenation are done in BAL subroutines to avoid inefficient PL/1 object code.

File Name	Use in FORCON
UFDBvvii	The data base tape (vv = vol, ii = issue). Kept permanently.
IFDBvvii	The IITRI-format tape (vv = vol, ii = issue). Kept 1 year.
DBOCPPRT	The FORCON (and, later OCP) listing. Destined for microfilm conversion.

5.3.3 IFCOPY

IFCOPY (IITRI Format COPYing program) takes an IITRI-formatted tape and copies it. It is used to produce a single file from each volume of a data base. This file can then be used for retrospective searches. Approximately 85,000 IITRI-format citations (without abstracts) fit on one tape reel.

File Name	Use in IFCOPY
IFDBvvii	The input IITRI-format tape file (vv = vol, ii = issue). Kept 1 year.
IFRFDBvv	The IITRI-format retrospective file (vv = volume); new records are appended. Kept permanently.

5.3.4 DKEDIT

DKEDIT (Profile Deck EDITor) scans search profiles for errors. Cards are checked for internal errors and the profile as a whole is checked to verify the information in the header card. The logic statement is checked for consistency with the terms and links read. Search terms that would match any of the 50 most frequent terms in CA are flagged.

File Name	Use in DKEDIT
Cards	The keypunched profile cards to be checked (see the Data Set Description for PCSCCARD for the structure of this file).

5.3.5 MINIPUP

MINIPUP (MINI-Profile Update Program) merges a profile update deck into another profile deck. In practice it is used to merge changes into a permanent profile stream stored on tape. A special card is used to drop profiles without replacement. An intermediate data set used in the process is stored on tape and can be used to re-create the output tape if it should be lost through machine malfunction during processing.

File Name	Use in MINIPUP
Cards	The keypunched cards containing profiles to be inserted (see the Data Set Description for PCSCCARD for the structure of this file).
PCSCCARD	The existing profile stream, into which the updates are inserted. This file is used both for input and output.
PCSCBKUP	An intermediate work file which contains all information necessary to create the output version of PCSCCARD.

5.3.6 INPUTR

INPUTR (Profile INPUT and Reformating program) reads the profile stream and builds the data structures used to describe the profiles in SEARCH. Terms are aggregated and divided into groups by Least Common Bigram. Logic expressions are expanded, by replacing links with the disjunction of their component terms, and converted to Early Operator Reverse Polish form. Profile information blocks are constructed from the header cards and other information.

File Name	Use in INPUTR
PCSCCARD	The profile stream to be converted.
PCSCTEXT	The index to the aggregated term list.
DBLCB	LCB's for the specific data base.
PCSCTERM	Unique search terms, sorted on LCB.
PCSCHEAD	Profile information blocks.
PCSCLOGC	Logic expressions for all profiles in the run.
PCSCPASS	The run communications file, used to pass data to SEARCH, STIXA, etc.

5.3.7 SEARCH

SEARCH reads the profile description structures created by INPUTR and uses them to search an IITRI-format tape. The search proceeds by reading one citation at a time and determining for which (if any) profiles the citation is a hit. If the citation was a hit, one copy of the citation is written to the hit file for each profile for which it was a hit. After all citations have been read a file is written containing the information necessary to build the Search Term Frequency/Issue listing which accompanies the citation cards.

File Name	Use in SEARCH
IFDBvvii	The IITRI formatted tape to be searched (vv = vol, ii = issue). Kept 1 year.
PCSCTEXT	The index to the aggregated term list.
PCSCTERM	The aggregated term list--all terms contained in all profiles in the run with duplicate terms removed, in order on LCB.
PCSCLOGC	The profile logic expressions.
PCSCPASS	The system communication file containing data passed from INPUTR and used to pass data to OCP and STIXA.
PCSCCITS	The citations retrieved (each citation that was a hit is included once only, regardless of how many profiles found it).
PCSCHITS	The hits found--one record for each citation found for each profile.
PCSCSTFD	The data needed to produce the Search Term Frequency/Issue listing.

5.3.8 DBCARD

DBCARD (Data Base CARD formatting program) reads the file of citations retrieved and builds, for each citation, a card image. The format of the card is determined, within limits, by the sizes of the various fields of the citation.

File Name	Use in DBCARD
PCSCCITS	The unique citations file produced by SEARCH.
PCSCFMCT	The output card images for the citations that were hits.

5.3.9 DB-OCP (Data Base Output Control Program)

5.3.9.1 ØCP-ØCP1

ØCP-ØCP1 (Output Control Program, Step 1) makes multiple copies of the citation card images, one for each hit recorded for each citation.

File Name	Use in ØCP-ØCP1
PCSCHITS	The file of hits written by SEARCH.
PCSCFMCT	The citation card images written by DBCARD.
POUTPRNT	The expanded hit-citation file; each record is a complete description of a hit, including the citation card image.

5.3.9.2 ØCP-SØRT 1

ØCP-SØRT 1 sorts the profile information blocks into profile number order.

File Name	Use in ØCP-SØRT 1
PCSCHEAD	The profile information blocks written by INPUTR.
PGSCSTHD	The sorted profile information blocks.

5.3.9.3 ØCP-SØRT 2

ØCP-SØRT 2 sorts the records in the expanded hit-citation file, written in ØCP-ØCP1, into citation number order within each profile. It also applies any special sorts requested for output.

File Name	Use in ØCP-SØRT 2
POUTPRNT	The expanded hit-citation file written by ØCP-ØCP1.
PCSCSRTD	The sorted expanded hit-citation file.

5.3.9.4 ØCP-ØCP2

ØCP-ØCP2 (Output Control Program, Step 2) reads the hit-citation file; it inserts into citations the weights and search terms found, creates header and trailer cards, applies print limits, builds Search Term Frequency/Issue cards, and writes the file of card images that produces the output cards and a compressed listing, without blank lines, for COM output.

File Name	Use in ØCP-ØCP2
PCSCSTFD	The data used to create the Search Term Occurrences listings.
PCSCRTD	The (sorted) expanded hit-citation list.
PCSCTHD	The (sorted) profile information blocks.
PCSCPASS	The system communication file.
PCSCØPLG	The printout counts for each profile, used by STIXA.
PCSCEXTR	A file of card images for tape output.

POUTPRNT	The print file of card images. This file is printed on 5" x 8" cards by an IBM utility program or off-line printing unit.
PCSCPRNT	A file used to hold header and trailer card images temporarily.
DBOCPPRT	The COM listing. The OCP listing is placed after the FORCON listing on this tape.

5.3.9.5 OCP-MICRO

OCP-MICRO translates the FORCON-OCP listing into a form suitable for use with the specific COM unit used to produce a microfilm copy of the listings.

File Name	Use in OCP-MICRO
DBOCPPRT	The FORCON-OCP listing.
MICRO.OUTPUT	The microfilm-format tape.

5.3.10 HITTER

HITTER generates a list showing, for each citation found by each profile, the search terms found in the citation.

File Name	Use in HITTER
PCSCHITS	The hit list file written by SEARCH.
PCSCSTHT	The hit list file, sorted on profile number.

5.3.11 STIXA

STIXA produces a statistical summary for each issue's run. Included are breakdowns of hits and prints by profile as well as the sizes of various files, as provided by the creating program. (See Production Statistics, Section 10.4 of this report.)

File Name	Use by STIXA
PCSCPASS	The system communication file, contains various statistics supplied by individual programs.
PCSCOPLG	A file of hit and print counts per profile.

5.3.12 PLSXT

PLSXT (Private Libraries System Extraction program) extracts the output from profiles for which Private Libraries System (PLS) files are to be created, converts the output to PLS format, and merges the result into PLS Master File. A copy of the previous Master File is made, to function as a back-up to the updated Master File.

File Name	Use in PLSXT
PCSCHITS	The hit file.
PCSCCITS	The unique citations retrieved file.
PCSCHEAD	The profile information block file.
PRIMASTR	The updated PLS Master File.
PRIWKOUT	The new citations that were added to the PLS Master File.
PRISRTOT	The old PLS Master File (created as an emergency back-up file).

5.4 Core Storage Requirements

All of the Computer Search Center programs can be run with a minimal configuration containing 256K bytes of core storage, two tape drives and one disk. However, modifications of the files can change these requirements to some extent. Our current sizes are designed for compute-bound operation on an IBM 360/65 computer, using 300K bytes of core for the largest (SEARCH) program. Core requirements can be decreased with a corresponding increase in I/O time for a smaller computer. If the smaller computer were also slower (e.g., 360/50) more concurrent time would be available for I/O resulting in no overall decrease in efficiency. The program requirements in the current operating environment are:

CSC Programs

DKEDIT	<95K bytes 1 sequential file (disk)
FORCON	<110K bytes 2 sequential files (tape)
INPUTR	(180K)+(13 x no. of profiles) + (31 x no. of terms) bytes 1 sequential file (tape) 7 sequential files (disk) 1 direct-access file (disk) 9 IBM SORT/MERGE files (disk)
SEARCH	(150K)+(547 x no. of profiles)+ (30 x no. of terms)+((no. of profiles + 1) x ([no. of terms/8]+1) bytes (note: using number of unique terms) 1 direct-access file (disk) 1 sequential file (tape) 4 sequential files (disk)
DBCARD	<110K bytes 2 sequential files (tape)
STIXA	<85K bytes 1 direct-access file (disk) 1 sequential file (disk)
ØCP-ØCP1	<80K bytes 1 sequential file (disk) 2 sequential files (tape)

CSC Programs (continued)

OCP-SORT1	128K bytes 2 sequential files (disk) 4 SORT/MERGE work files (disk)
OCP-SORT2	200K bytes 2 sequential files (tape) 16 SORT/MERGE work files (disk)
OCP-OCP2	<115K bytes 1 direct-access file (disk) 2 sequential files (disk) 5 sequential files (tape or disk)
OCP-MICRO	<55K bytes 2 sequential files (tape)
PCSXT	≤150K bytes 2 sequential files (disk) 5 sequential files (tape or disk) 6 SORT/MERGE work files (disk)
IFCOPY	<65K bytes 2 sequential files (tape)
MINIPUP	<85K bytes 4 sequential files (tape or disk)
DBCOPY	<65K bytes 2 sequential files (disk)

IBM Utilities

IEBPTPCH	60K bytes 1 sequential file (tape)

5.5 System Files and File Structures

The various programs that constitute the SDI system communicate via a system of files. These files are used to pass blocks of data created at each step to the step or steps that use them later or to permanently hold data for the archives. The data in each file have a carefully-defined structure that is used in reading or writing the file. In the following file descriptions we have defined the data structures in terms of the PL/1 data declarations used to read or write the file. For details of the resulting physical data arrangement, see the IBM Systems Reference Library publication GC28-8201, PL/1(F) Language Reference Manual.

5.5.1 CA-CØND.SDF1 (This is an example of a data base tape.)

Description:

This tape is the CA Condensates data base tape in Standard Distribution Format (SDF) as supplied by CAS. It typically contains 5000-8000 citations.

Format:

For a description of Standard Distribution Format see the following CAS publications:

Standard Distribution Format, Technical Specifications (Revised); Condensates in S.D.F., Data Content Specifications

History:

Creation:	Supplied by CAS
Referenced:	Read by CACOPY
Disposition:	Scratched after copying

5.5.2 DBOCPPRT (Data Base OCP Print tape)

Description:

This tape contains the FORCON (see Section 5.3.2 of this report) and OCP (see Section 5.3.9 of this report) listings and is used to create a microfilm copy of those listings.

Format:

Block of 50 records, each 121 characters long. Each record is a line image written to a PL/1 PRINT file.

History:

Creation:	FORCON
Update:	Read by MICRO (the OCP microfilm step)
Disposition:	Held one week, then re-used in same capacity

5.5.3 UFDBvvii (Unformated Data Base copy, volume vv, Issue ii)

Description:

This tape is a copy of the data base tape in supplier format for volume vv, issue ii.

Format:

This tape is in the supplier's format. For details see the documentation for the specific data base. Number of issues per copy reel will vary with data base.

History:

Creation:	Written by DBCOPY
Referenced:	Read by FORCON
Disposition:	Kept permanently

5.5.4 IFDBvvii (IITRI Format Data Base tape, volume vv Issue ii)

Description:

This tape is an IITRI-format equivalent of UFDBvvii, where vv = volume number, ii = issue number. This is the SDI search tape.

Format:

Varying length records up to 4000 characters long in blocks up to 4000 characters long. For a complete description of IITRI format, see Section 5.2 of this report.

History:

Creation: Written by FORCON (six issues per reel)

Referenced: Read by IFCOPY

Read by SEARCH

Disposition: Kept one year

5.5.5 IFRFDBvv (IITRI Format Retro File Data Base, volume vv)

Description:

This tape is a retrospective data base containing IITRI formatted tapes for a given volume. Typically a volume takes slightly more than two reels, though this varies according to data base.

Format:

Varying length records up to 4000 characters long in blocks up to 4000 characters long. For a complete description of IITRI format, see Section 5.2 of this report.

History:

Creation Written by IFCOPY (DISP-MOD)

Disposition: Kept permanently

5.5.6 M.CRO.OUTPUT (Microfilm Output)

Description:

This is a tape formatted for input to a COM system for production of microfilm output. It contains the FORCON and OCP listings for an issue's run.

Format:

This is an 800-bpi tape. The format is specific to the COM unit used.

History:

Creation: Written by 0CP2

Disposition: Re-used for next issue after production of microfilm.

5.5.7 PCSCBKUP (PCSCCARD Backup)

Description:

This is an intermediate file used in MINIPUP (see Section 5.3.5 of this report). It contains the results of the merge, but not the run header card. It provides a back-up to the output tape which can be retrieved by MINIPUP using a control card.

Format:

Blocks contain 80 fixed-length 80-character records. The order and delimiters are the same as for PCSCCARD except there is no run header card.

History:

Creation:	Written by MINIPUP
Referenced:	Read by MINIPUP
Disposition:	Saved until next issue, then re-used

5.5.8 PCSCCARD (Profile Card stream)

Description:

This tape contains the profile stream for input to INPUTR.

Format:

Blocks contain 80 fixed-length 80-character records. Each is the image of a punched card.

The deck consists of:

(1) Run header card = 'PPPTTTTT' where PPP = number of profiles, TTTTT = number of terms

(2) Profile header and term cards, ordered on positions 3-9

(3) Delimiter card = 'DE-LIMITER000000000LAST TERM CARD'

(4) Logic expression cards, ordered on positions 3-9

(5) Delimiter card = 'DELIMITERƀƀƀƀLAST LOGIC CARD'

For details of header, term, and logic formats, see Sections 5.5.11, 5.5.13, and 5.5.19 of this report.

History:

Creation:	Written by IEBGENER from cards
Update:	Read and rewritten by MINIPUP
Referenced:	Read by INPUTR
Disposition:	Updated and used for each issue (in the case of CA Condensates, every other issue; we use separate tapes for CA even and CA odd issues)

5.5.9 PCSCCITS (Unique Citations retrieved file)

Description:

This file contains one copy of each citation retrieved by SEARCH.

Format:

Blocks of two fixed-length 1251-character records.

Records are written and read with the PL/1 structure:

```
1 CITATION_RECORD,  /* contains one citation from the data*/
                    /* base, in IITRI-format              */
    2 ABSTRACT_NUMBER character (11),
                    /* the unique abstract no., 11 digits */
    2 DIRECTORY (60) fixed binary (31),
                    /* the IITRI-format directory;        */
                    /* contains sixty fullword binary     */
                    /* numbers                            */
    2 STRING character (1000),
                    /* the IITRI-format string part       */
                    /* a fixed-length version for better  */
                    /* processing efficiency              */
```

History:

Creation: Written by SEARCH

Referenced: Read by CACARD

Read by ØCP1

Read by PRILIB

Disposition: Re-used for each issue

5.5.10 PCSCFMCT (Formatted Citation file)

Description:

This file consists of images of 5" x 8" printout cards. These are generated by DBCARD from the citations in PCSCCITS.

Format:

Blocks contain two fixed-length 2411-character records. Each record consists of an 11-character citation number and 38-character line images.

History:

Creation: Written by CACARD

Referenced: Read by ØCP1

Disposition: Re-used for each issue

5.5.11 PCSCHEAD (Header information file)

Description:

This file contains profile information extracted in INPUTR from the profile header card and the profile itself.

Format:

Blocks contain 100 fixed-length 26-character records.

Records are read and written with the PL/1 structure:

```
1 HEADER,              /* this is a HEADER information block*/
      2 PROFILE_NUMBER character (10),
                       /* the ten-character user i.d. number*/
      2 WEIGHT_THRESHOLD fixed decimal (5),
                       /* the minimum weight required for a */
                       /* citation to be retrieved          */
      2 OUTPUT_DEFINITION character (4),
                       /* position 1: output medium (C or P)*/
                       /* position 2: number of copies      */
                       /* positions 3-4: sort type for output*/
      2 PRINT_LIMIT fixed decimal (5),
                       /* maximum number of citations to be */
                       /* printed
      2 SORT_FIELD_LENGTH fixed decimal (5),
                       /* length of the field selected for  */
                       /* the output sort                   */
      2 EXTRACTION_REQUEST character (3),
                       /* contains 'PRI' if the profile's   */
                       /* output is placed in a private     */
                       /* library                           */
```

History:

Creation:	Written by INPUTR
Referenced:	Read by SEARCH
	Read by ØCP.SØRT1
	Read by PRILIB

5.5.12 PCSCHITS (Hits Recorded)

Description:

This file contains one record for each citation found for each profile. It is used to construct the output stream from the file of unique citations.

Format:

Blocks contain 20 fixed-length 148-character records.

Records are read and written with the PL/1 structure:

```
1 HIT_RECORD,          /* describes one hit (i.e. a single    */
                       /* citation matching a single profile) */
    2 PROFILE_NUMBER character (10),
                       /* the ten-letter user i.d. number     */
    2 HIT_WEIGHT fixed decimal (5),
                       /* the retrieval weight of the citation*/
                       /* for this profile                    */
    2 CITATION_NUMBER character (11),
                       /* the abstract number of this citation*/
    2 SORT_FIELD character (45),
                       /* the actual string that will be used */
                       /* as a sort key in ordering the output*/
    2 SEARCH_TERMS character (79),
                       /* a string containing the search terms*/
                       /* that were present in the citation   */
                       /* from this profile                   */
```

History:

Creation:	Written by SEARCH
Referenced:	Read by ØCP1
	Read by HITTER
	Read by PRILIB
Disposition:	Re-used for each issue

5.5.13 PCSCLOGC (Profile Logic Expressions)

Description:

This file contains the logic expressions for the search profiles. The expressions consist of terms--represented by term numbers in the Aggregated Term List--and operators and are in Early Operator Reverse Polish (EORP) form.

Format:

Blocks contain a single varying-length record with a maximum length of 4630 characters. The PL/i structure used for these is:

```
1 LOGIC_EXPRESSION, /* the internal representation of a    */
                    /* profile's logic expression          */
    2 THRESHOLD_TERMS fixed binary (31),
                    /* the minimum number of terms which   */
                    /* must be found for the logic to be   */
                    /* satisfied                           */
    2 BIT_ARRAY (EXPTOT) bit (1),
                    /* a vector containing one bit for     */
                    /* each term in the A.T.L.  The size   */
                    /* of the vector (EXPTOT) is read at   */
                    /* initialization time                 */
```

```
2 PROFILE_NUMBER character (10),
               /* the ten-letter user i.d. number    */
2 SORT_OPTION character (2),
               /* the output sort type; WT for a     */
               /* sort by weight, AN for abstract    */
               /* number order, 03 for author order  */
2 THRESHOLD_WEIGHT character (3),
               /* the minimum weight necessary for   */
               /* retrieval, a zoned-decimal number  */
2 NUMBER_OF_SYMBOLS character (3),
               /* the total number of terms and      */
               /* operators in the expanded logic    */
               /* expression, a zoned-decimal number */
2 EXPRESSION character (4200) varying,
               /* the actual terms and operators, in */
               /* seven-letter blocks containing the */
               /* term number (in the A.T.L.) and    */
               /* weight for terms, or operator code */
               /* for Boolean operators              */
```

<u>History</u>:

Creation: Written by INPUTR

Referenced: Read by SEARCH

Disposition: Re-used for each issue

5.5.14 PCSCOPLG (Output Logging File)

<u>Description</u>:

This file consists of accounting information for STIXA. One record is written for each profile to generate the breakdown of hits and prints by profile.

<u>Format</u>:

Blocks contain 200 fixed-length 16-character records.

Records are read and written with the PL/1 structure:

```
1 OUTPUT_LOG_RECORD,
                    /* this record contains the number of */
                    /* hits recorded for a given profile  */
    2 PROFILE_NUMBER character (10),
                    /* the ten-letter user i.d. number    */
    2 PRINT_LIMIT fixed decimal (5),
                    /* the maximum number of citations to */
                    /* be printed by this profile         */
    2 NUMBER_OF_HITS fixed decimal (5),
                    /* the number of citations retrieved  */
                    /* by this profile                    */
```

History:

Creation: Written by ØCP2
Referenced: Read by STIXA
Disposition: Re-used for each issue

5.5.15 PCSCPASS (Run Information Passing File)

Description:

This is the system communication file. It contains six values which are set and read at various points in the system.

Format:

This is a direct-access file with REGIONAL (1) organization. Each record is a fullword binary number. The entries are:

(1) Number of citations searched
(2) Number of profiles
(3) Number of terms
(4) Number of unique terms
(5) Number of unique citations retrieved
(6) Volume and issue (vvii, a 4-digit number)

History:

Creation: Permanently set-up
Updated: Written from INPUTR
Written from SEARCH
Written from ØCP2
Disposition: Re-used for each issue

5.5.16 PCSCSRTD (Sorted expanded hit-citation file)

Description:

This file is a sorted version of PØUTPRNT$_1$ (q.v.), the expanded hit-citation file. The sort is on profile number and output sort field, as selected by user.

Format:

Block contain 2 fixed-length 2548-character records. For detailed record format see: PØUTPRNT$_1$.

History:

Creation:	Written by ØCP-SØRT 2
Referenced:	Read by ØCP2
Disposition:	Re-used for each issue

5.5.17 PCSCSTHD (Sorted Header Information file)

Description:

This file is the same as PCSCHEAD (q.v.), but sorted on profile number.

Format:

Blocks contain 100 fixed-length 26 character records. For record format see: PCSCHEAD.

History:

Creation:	Written by ØCP-SØRT1
Referenced:	Read by ØCP2
Disposition:	Temporary disk file, released at end of run

5.5.18 PCSCSTHT (Sorted Hit file)

Description:

This file is a copy of PCSCHITS (q.v.), sorted on profile number. It is used by HITTER to create its listing of search terms found in citations retrieved.

Format:

Blocks contain 30 fixed-length, 148-character records. For record format, see: PCSCHITS

History:

Creation:	Written by sort in HITTER
Referenced:	Read by HITTER
Disposition:	Temporary disk file, released at end of run

5.5.19 PCSCSTFD (Search Term Frequency Data file)

Description:

This file contains the counts of the number of citations containing each search term in each profile. This data is used to generate the Search Term Occurrences cards in the output stream.

Format:

Blocks of 200 fixed-length, 13-character records.

Records are read and written with the PL/1 structure:

```
 1 STF_DATA,           /* an item in the Search Term Frequency */
      2 PROFILE_NUMBER character (10),
                       /* the ten-letter user i.d. number      */
      2 LINK_NAME character (1),
                       /* the letter used to designate the     */
                       /* link in which this term appeared     */
                       /* if unlinked, then contains 'ƀ'       */
      2 TERM character (22),
                       /* the search term                      */
                       /* asterisks are added at left and/or   */
                       /* right ends to show truncation mode   */
      2 COUNT fixed binary (31),
                       /* a fullword binary number giving the  */
                       /* a number of citations in which this  */
                       /* term was found                       */
```

History:

Creation:	Written by SEARCH
Referenced:	Read by 0CP2
Disposition:	Re-used for each issue

5.5.20 PCSCTERM (Aggregated Term List)

Description:

This file contains the aggregated term list and the information needed to match the terms against the citation search string.

Format:

The file consists of blocks of 20 27-character records, each of which consists of the search term, with the LCB prefixed to it, the offset of the LCB from the beginning of the term (so that the term can be 'slid' into the correct orientation on the citation string), and the truncation mode, expressed as a single character.

History:

Creation:	Written by INPUTR
Referenced:	Read by SEARCH
Disposition:	Re-used for each issue

5.5.21 P0UTPRNT$_1$ (Note: this name is used for two different files)

Description:

This file is an expanded hit-citation list, with the citation card image appended to each hit record.

Format:

Blocks contain 2 fixed-length 2548-character records. Record format is the same as for PCSCHITS except that a new element '2 CARD_IMAGE character (2400)' is added to the structure to contain the 30 80-character line images.

History:

Creation:	Written by 0CP1
Referenced:	Read by 0CP-S0RT2
Disposition:	Tape is re-used immediately by 0CP2

5.5.22 P0UTPRNT$_2$ (Note: this name is used for two different files.)

Description:

This is the print tape and consists of line images to be printed by an IBM utility or off-line printer whenever convenient.

Format:

Blocks contain 60 fixed-length 80-character records. Each record is an image of a line of output to be printed.

History:

Creation:	Written by 0CP2
Referenced:	Read by printing routine
Disposition:	Re-used for each issue

5.5.23 PRIMASTR (Private Libraries Master Tape)
PRISRTOT (Private Libraries Restart Old Tape)
PRIWKOUT (Private Libraries Update Work Output Tape)

Description:

These three tapes are the Private Libraries Master tape, a back-up copy containing the previous Master, and an update tape containing the records that were added to the previous Master to produce the current Master.

Format:

These tapes contain variable-length records up to 4000 characters long in blocks up to 4000 characters long. For detailed format information, see the Private Libraries System section of this report.

History:

Creation: Written by PLSXT
Update:
Referenced: By Private Libraries System (q.v.) as necessary
Disposition: Updated after each SDI system run

5.6 Search Algorithm Comparison

5.6.1 Introduction

A bibliographic search program must perform several functions. The first of these is matching search terms to citations. Since the CSC Software System is essentially an SDI search system, there are several constraints on this matching process. For SDI, one assumes a one-time use of the data base, so extensive reformatting is not cost-effective. Also, we would expect to search many profiles against each citation without re-reading the citations for each profile. Finally, for data bases with uncontrolled vocabulary (the class we are considering) it is necessary to: 1) check the types of information in the citation to limit search for specific data elements to the appropriate portion of a citation, e.g., to find author terms the search should be restricted to the author portion of a citation, etc., and 2) search on word fragments, words, multiword terms, and phrases.

If the citation can be divided into terms defined by readily recognizable delimiters, for example, words bounded by spaces, a number of variant search orders are possible. In one variant, each term in each profile is matched against each term in each bibliographic record. In a second variant, the profile terms can be inverted and each term of a bibliographic record is sequentially matched against all terms in the inverted profile term list. The profile term list is passed on as many times as there are terms in the bibliographic records.

In a third variant, profiles are inverted and each profile term is matched against each bibliographic record. The bibliographic record list is passed as many times as there are terms in the inverted profile term list. A fourth variant is based upon inverting the bibliographic records. This inverted file can be matched against profile terms, either profile by profile or in an inverted profile list.

Each of the above variant search procedures assumes that designated delimiters effectively distinguish terms. As we will discuss below, this is not wholly true in a complex, scientific data base. Term designation precludes phrase searching unless individual terms can be concatenated by means of a logic operator. An inverted bibliographic term list precludes left truncation.

The major problem with all of these methods that depend upon division of a citation into words is that of defining what is to be considered a word. The obvious supposition is that a word is any group of letters bounded by blanks. The other extreme is that all nonalphanumeric characters be considered delimiters. Neither of these options is an acceptable solution because in the first case, going from blank to blank makes all punctuation part of the preceding word (which is sometimes the case, in fact) and in the second, the fact that special characters can be parts of words is totally ignored. Implementing either of these alternatives would make effective profile preparation improbable since in the first case all words would have to be listed as truncated on the right to allow for punctuation problems (not to mention the problems arising with left parentheses) and in the second case, even simple abbreviations or hyphenated words would have to be entered into the system by individual letters and fragments. Although a compromise between the two solutions could be programmed, the resulting programs would not be general and would be open to continuous revision inasmuch as the input is essentially uncontrolled.

It then becomes apparent that in our SDI system for searching data bases with uncontrolled vocabulary, it is necessary to avoid arbitrary definitions of a term such as "characters delimited by blanks". Thus we must define a search term as an arbitrary string of characters to be matched against any given portion or portions of a citation. Thus the citation cannot be divided into terms but is treated as a string of characters and cannot be inverted. The functions of a search program then can be listed as:

- match search terms against the citation string
- check term types
- maintain assignment of potential hit terms to
- respective profiles in the batch
- evaluate profile logic expressions
- prepare hit records for subsequent processing

In this section we deal primarily with the first of these functions, the basic problem of matching search terms to citations and coming up with a "yes" or "no" answer for the presence of any given search term in any citation. However, the method in which this basic task is performed will affect some of the other search functions, and we will discuss them as they occur. We are assuming that, for reasons of one-time use of the data base and presence of uncontrolled vocabulary, the data base is to be searched serially, one citation at a time, for all the profiles in the system. To allow comparisons among the various methodologies that we discuss below, we will assume that there are 3000 terms in the search term list (200 profiles averaging 15 search terms each), 5000 citations on the data base issue being searched, and a citation length of 200 characters. These assumptions represent a typical SDI run for a data base such as EI COMPENDEX or CA Condensates.

In the following sections we discuss the various algorithms we have used to develop an efficient method of matching search terms against citations. The algorithms are presented in order of increasing efficiency. This is also the chronological order in which the various methodologies were used by the Computer Search Center.

5.6.2 Term-to-Citation Algorithm

The simplest method is to match each term in the search term list against each citation, letter by letter. The search terms are, of course, sequenced by type, so that author terms are matched only against the author portion of the citation, etc. In this method, when working on the title of a citation, all title terms in the search term list are checked against the entire title. In effect, this means that the program must take the first title search term and check if it matches the title, beginning at the first letter of the title. If not, the search term is "slid over" one character in the title and checked again. This is repeated until the last character of the title is reached. Then the whole process is repeated for the next title-type search term, etc.

This is a very simple and straightforward method. It can be coded very easily, In fact, in PL/1 there is a built-in function (INDEX) that permits a one-line coding of this method. However, it is extremely expensive. For our assumed SDI parameters, there would be 3000 times 5000 times 200, or 3 billion matches required. Although tested for information purposes, this highly inefficient algorithm was never used in production.

5.6.3 Basic Citation-to-Term Algorithm

A preliminary analysis of the problem indicates that the search should be done from citations to search terms rather than the reverse, since there are only some 200 characters in the citation and 3000 initial characters in the search term list. (Throughout this entire discussion, we are ignoring match on the rest of the characters of a search term after the entry character to the citation string has been found. This additional character matching is the same for all methodologies and so can be eliminated from the discussion.)

In this basic method, the program goes through the citation one character at a time and checks for match only those search terms that begin with the character then being considered. (The previous division by term type is assumed here also, as it is in the balance of the discussion.) The reason that this methodology is more efficient than the previous one is based on the fact that the search term list is divided into groups separated by initial letters. When an "A" is found in the citation, only those terms beginning with "A" are checked for match, rather than all the terms. Since we are working with a character set of 50 characters (alphabetics, numerics, and punctuation symbols), in the ideal case our 3000 search term list would be divided into 50 groups of 60 terms each. The number of matches would then be 60 times 5000 times 200 or 60 million. However, there are certain characters that are seldom found at the beginning of a word, for example, very few words begin with a semi-colon, thus in reality the average size of a group is 100 terms rather than 60, so that 100 million matches are required.

This methodology reduces the number of matches to 3.33% of those required by the first algorithm. It cannot be coded quite as simply, since tables must be maintained to point to the position in the search term list of each of the groups delimited by a different character and to indicate the number of search terms in each of the groups. However, building these tables is quite simple and using them does not add much to the cost of matching. Truncation is as easily checked as in the first case by looking at the character preceding and the character following the term after a match has been found. We used this algorithm as our first production methodology.

5.6.4 Basic Citation-to-Term via Initial Bigrams

Since the search should proceed from citations to search terms, our aim was to reduce the size of an average group of search terms and increase the number of groups. This would mean fewer matches for each locator in the citation. However, it must be accomplished without overly complicating the process of getting from the citation to the proper place in the term list. In the case above, this was accomplished by an alphabetic grouping of the search terms and a simple table, giving a very efficient route.

Since the first character division method gave us about 100 terms in the average-sized group, we next tried the initial two letters, or initial bigram. Theoretically, the terms would be divided in 2500 groups of 1.2 terms each, since there are 50-squared possible bigrams. But, as unlikely as it is to find a word beginning with a semi-colon it is even less likely to find one beginning with two semi-colons. In practice, we found that half of all the search terms fell into groups headed by one of 60 bigrams, and that the actual average group size was 20, not the ideal 1.2. This was still a great improvement, however, reducing matches to 20 x 5000 x 200 or 20 million. This is 0.67% or the first case and better than the initial letter method by a factor of five.

Implementing this algorithm requires a bit more coding. Skipping through the character string two characters at a time is not too difficult since the first letter of the second bigram was already found as the second letter of the first bigram, etc. The tables to the positions of the terms in the list that begin with each bigram are a bit more complex since they must be based on two values, one for each chacter of the bigram.

Prior to beginning the search, a bigram table of 2500 sets of two numbers is set up. A unique value for the bigram is determined by the formula $S = 50 P_1 - P_2$, where P_1 and P_2 are the positions of the first and second characters, respectively, in the 50-character string. A character string of length 50 is set up containing these 50 characters as ordered by frequency of appearance of single characters in the data base. The table is filled for each bigram; the first number of each set being the starting position in the term list of terms sorted on that bigram and the second number being the number of terms sorted on that bigram. For example, CHRSTR (1) refers to the letter E, and CHRSTR (9) to the letter H. The value of the bigram EH, is thus 50 × (1) - (9) or 41. If words sorted on EH were the 189th through 197th terms in the term list, TABLE (41,1) would be 189 and TABLE (41,2) would be 9. Filling this table is very rapid and is done only once at the beginning of the program. Table positions for which no terms exist in the term list are set to -1. As the search proceeds through the citation, two letters at a time, a check is made in TABLE for each value calculated. If a a -1 is found, no term check is made. If a positive value is found for TABLE (N,1) matches are made starting with the value in TABLE (N,1) and continuing for the number contained in TABLE (N,2). All in all, the overhead was not very much higher in terms of either machine time or storage requirements, and so we put this algorithm into production at a considerable increase in efficiency.

5.6.5 Basic Citation-to-Term via Initial Trigram

With the experience of initial letter and bigram matching at hand, the next logical extension was to trigrams (three letters). The same procedures could be followed but would be extended to three-character sets. In theory, there would be 125,000 (50^3) possible groups into which the terms could be divided, but since we only have 3000 total search terms in the list, the best that could be achieved would be 3000 groups with one term in each. In practice, this would not quite be realized, since there are some very common trigrams in English, such as "THR", "CRE", etc. The average group size would be about 1.5. This would result in 1.5 times 5000 times 200 or 1.5 million matches, and would be much better than the bigram method. It looked worthy of implementation, but the overhead required to maintain the locator tables was too high for trigrams. The size of the tables grows exponentially. In the initial letter case it had 50 entries, for bigrams it had 50 squared entries, and for trigrams it had 50-cubed entries. The first two tables fit easily in core storage (100 and 5000 bytes, respectively) but the third needs 25 million bytes and would have to be compressed to fit in core. The compression and coding necessary to decompress each time the table was entered (for each character of each citation) made the overhead much higher than the savings realized by using trigrams, and so this algorithm was not implemented. We estimate that the trigram method would begin to show improved efficiency over the bigram method for term lists containing 50,000 to 100,000 items.

5.6.6 Basic Citation-to-Term via Least Common Bigram

Having determined that searching from citations to search terms via initial bigram lookup was efficient, we used this method in production for more than a year. However, we had never given serious thought as to why we used the initial bigram as the locator. Since dictionaries are commonly ordered alphabetically on letters from left to right, we took this as a natural way to order a word list. In point of fact, selection of any bigram but the first one would have divided the search term list into more groups, with fewer terms per group, on the average. Having noticed this, it was simple to determine which bigram should be chosen for any given search term. That bigram is the Least Common Bigram (LCB), i.e., that bigram that appears least frequently in the data base. Since we had prepared KLIC Indexes, we knew the frequency distribution for all the bigrams in each data base. Using the LCB screen technique as search terms enter the system, a small routine checks them against the bigram frequency table for the appropriate data base. For example, the word "MOLYBDENUM" contains the bigrams MO, OL, LY, YB, DE, EN, NU, and UM. The routine checks each of these in the table and finds the LCB, the bigram with the lowest frequency, in this case, BD. MOLYBDENUM is then maintained in the search term list under the bigram (LCB) BD rather than the initial bigram MO. The TABLE for finding the proper position in the search term list is maintained and used exactly as it was for the basic bigram technique.

This technique of using LCB's provides two improvements in efficiency. First, more bigrams are used within words than are used to begin words. For example, no words begin with "KK", yet "BOOKKEEPING" and other words contain it. Thus words are divided into more groups, with fewer words per group. In practice, groups average 5 terms in size, making the number of matches 5 times 500 times 200 or 5 million. This is only one-fourth as many as for initial bigrams. The second beneficial feature of the use of LCB's is that the largest groups are sorted under LCB's that occur least frequently, so the largest groups are searched less often than the smaller ones. These two factors combine to make an LCB-based algorithm highly efficient.

The additional machine time to arrange terms by LCB is very minimal, and little extra coding is required to search on this basis. It is necessary to maintain a number for each word that indicates how many characters from the beginning of the word the LCB is located, so that proper screening checks can be made. An example of search using the LCB technique follows.

Once the LCB screen has indicated the portion of the search term list to be searched, each term in the relevant profile term sublist is compared with the section of text indicated by the screen LCB as a possible match for that term. The relevant portion of text is delimited by referring to a number pair associated with each term. The first number tells where the compare area begins in relation to the screen LCB under consideration, and the second number gives the length of the character string to be compared. For example, suppose a title that includes the phrase ...PRESENCE OF ALDEHYDES IN ... is being searched against the term list. Taking each bigram in turn as an entry to the list the search algorithm will come in due course to the bigram EH and access from TABLE the information that entries referenced by the LCB EH start at term 526, and that there are nine of them:

Term	LCB	Term Type	Term	Required Truncation Mode	Characters Preceding LCB	Term Length
526	EH	02	ACETALDEHYDE	1000	7	12
527	EH	02	ALDEHYDE	0010	3	8
528	EH	02	ALDEHYDE OIL	0010	3	12
529	EH	02	BUTYRALDEHYDE	1000	8	12
530	EH	02	DEHYDROGENASE	0011	1	13
531	EH	02	DEHYDROGENAT	0010	1	12
532	EH	02	FORMALDEHYDE	1000	7	12
533	EH	02	PROPIONALDEHYD	1000	10	14
534	EH	02	VALERALDEHYDE	1000	8	13

Present in the core image of the term list are the additional numbers written in the sample above. Starting with term 526, and using the last two numbers 7 and 12, the search routine delimits the compare area in the title as follows:

7 characters screen
preceding bigram

. . .P R E S E N C E O F A L D E H Y D E S I N . . .

total term length
12 characters

and performs a compare on the character strings ACETALDEHYDE and ƀOFƀALDEHYDE. The result is not an equality, so the search goes on to term 527 and delimits the title again:

3 characters screen
preceding bigram

. . .P R E S E N C E O F A L D E H Y D E S I N . . .

total term length
8 characters

This time it compares the term character string ALDEHYDE and the text character string ALDEHYDE. This compare indicates the existence of a match, so the program goes on to determine the truncation modes that are satisfied. Testing the positions on either side of the compare area, the program finds a non-alphanumeric character (a blank) preceding the term and an alphanumeric character S following the term. Thus this citation satisfies the requirements for "right" and "both" truncation modes, and its found truncation mode is 0011. Combining this with the required truncation mode (0010) by a logical AND operation gives a nonzero result (0010). Thus term 527 is a hit term. The additional terms referenced by the LCB EH are then tested by the process outlined above, with no more hits resulting.

The immediately obvious advantage of the LCB search method is that only a subset of the term list is checked in each match. Further, since there are only 50 characters in the set, although the term list increases in size, the number of sets of two numbers remains the same. It then follows that a two-fold increase in the size of the term list will not result in a two-fold increase in search time. Thus, the rate of increase of search time decreases as a function of increasing number of terms. This technique shows a time savings for more than 120 terms.

5.6.7 Summary

The Least Common Bigram algorithm is based on a number of discrete steps, each of which gave more insight into search algorithms and increased efficiency. It is a very good algorithm and is based on the characteristics of the data base being searched. The table below summarizes the evolution of our search methodology.

METHOD	MATCHES/ISSUE*
Terms vs Test	3,000,000,000
Initial Letter	100,000,000
Initial Bigram	20,000,000
Initial Trigram	1,500,000**
Least Common Bigram	< 5,000,000

*Based on 3000 search terms and 5,000 200-character citations

**Increase in overhead for processing more than negated savings.

5.7 Logic Evaluation

The CSC system allows the use of the Boolean operators "AND," "OR," and "NOT" nested to any degree to indicate the relationship of search terms in a profile. The relationship of the terms in this way is called a logic expression and consists of term numbers (or link characters for those terms grouped in a link), the operators and parentheses to indicate the order in which the operators are to operate upon the term representations (operands). The profile writer is free to use as many parentheses as necessary to express the concepts imbedded in the term relationships.

When one or more of the search terms in a given profile is found in a citation, the logic expression for that profile must be evaluated to determine if a "true" hit has been found. To facilitate machine evaluation of logic expressions, they are converted, at profile input time, from the parenthetical notation used by the profile writer, to an unambiguous parenthesis-free notation. One such form of notation is called Polish notation, after the nationality of its inventor.

5.7.1 Early Operator Reverse Polish

CSC uses the Early Operator Reverse version of this notation, commonly called by its acronym, EORP. There are also "Late" versions and "Forward" versions, giving a total of four combinations, EORP, EOFP, LORP, and LOFP. The "Late" and "Early" refer to the relative positions of operators and operands, while the "Reverse" and "Forward" refer to the direction of evaluation of the expression.

EORP notation is based on assignment of preference to the operators. Thus a program can be written to convert parenthetical notation to this form, by replacing parentheses that denote operational order with one based on operator preference. Consider the two simple expressions:

(A & B) | C

(A | B) & C

(where, & = AND, | = OR, ¬ = NOT) which are clearly not identical. In EORP, these would be respectively:

AB & C |

AB | C &

EORP expressions are evaluated by proceeding from left to right, performing the operation called for by each operator upon the preceding two elements (except for NOT which is a unary operator). The result of each such operation is an operand in the next stage. To indicate the sequence of an evaluation, consider the following expression as an example:

(((A | B) & (C | D)) & E) & ¬F

which becomes, in EORP:

AB | CD | & E & F¬ &

To show the evaluation, we assume that A, B, D, and E are present (we will use "1" for present or True and "0" for not present or False). The expression is evaluated as shown in the steps below:

```
11 | 01 | & 1 & 0¬ &     Expression with '1' and '0'
 1   01 | & 1 & 0¬ &     Evaluation of 1st Operator
 1    1   & 1 & 0¬ &         "      "  2nd    "
   1        1 & 0¬ &         "      "  3rd    "
     1          0¬ &         "      "  4th    "
     1           1 &         "      "  5th    "
          1                  "      "  6th    "
```

The result is True. In each line the next operator is evaluated, the result "dropped down," and the operator is removed. The process is continued until all operators have been checked and a True or False answer results.

The major failing with EORP notation evaluation is that the entire expression must be checked before the final result is known. Consider the expression:

A & (B | C | D | E | F | G | H | I | J)

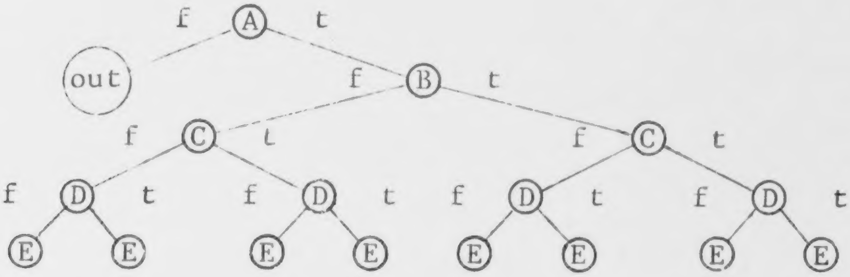
which in EORP would be

ABC | D | E | F | G | H | I | J | &

Only when the last & is checked is the result found. Yet it is immediately obvious, in the parenthetical notation, that is A is not present, the expression is definitely False. To get around this drawback of EORP notations, we have been considering and testing two alternative logic evaluation systems.

5.7.2 Tree Logic

An alternative to the EORP logic system would be generating a tree to represent the logic expression. Each operator would be a node. The expression above would generate the tree:



etc.

If A is not True, evaluation ceases immediately. If all subsequent false branches are followed, the result is False. If any subsequent True branch is followed, the result is True. It is necessary to follow the whole tree down rather than exiting as True if A and any of the others is found to enable detection of all True others. On the average, this type of evaluation should allow finishing evaluation in half the time required for EORP notation evaluation.

However, constructing the trees is difficult especially for recursive expressions and those that use the same term more than once. We are still testing this technique.

7.3 Modified EORP

A second alternate logic system would involve retaining the EORP notation, but for terms grouped together in a link (all terms in a link are implicitly OR'd together), a second expression would be generated. An initial evaluation would be made of the short expression (the one with only one operand per link). Only if the expression were found to be true would the entire expression be checked (the one with an operand for each term in each link). This system appears simpler to implement and we are now running timing tests on it.

5.8 Private Libraries System

5.8.1 Private Libraries System--General

The Private Libraries System is a software system that is used for the creation, maintenance, and searching of private files or subset data bases in machine-readable form. A private library can be established for an individual, a laboratory, a company, or any other organizational unit. Input to a private library can be from any source specified by the requestor. It may be citations, abstracts, full text, or document surrogates containing virtually any kind of data element the user wishes to retain. The documents may be company reports, literature, references, laboratory log books, correspondence files, etc. The data elements may be authors, titles, project numbers, key words, index terms specified by file users, codes corresponding to any meaningful data parameter the user may wish to record, etc.

The user who has a private file established and maintained for him controls the input to the file. He determines what should go into the file, what should be deleted from the file, and when the file should be purged. All of the items in the file represent his judgments and decisions as to relevance. He may wish to have his weekly output from the SDI system automatically entered into his private library for use at a later date or he may want to look at the output to determine which citations should be included and which should not. Additionally, he may want to enrich the citations by adding indexing terms, codes, or categories that have meaning to him or his company such as project numbers, product numbers, etc. The net result is a personally tailored file in machine-readable form that the user may search on demand.

The Private Libraries System can be adapted to accomodate virtually any existing file of document related data. Hence, a company that wants to establish a computerized retrieval system for its files need not go to the expense of time-consuming and costly software design and development it can have its files converted to IITRI format and then use an existing software system.

5.8.2 Software System

The Private Libraries System (PLS) is a group of programs which interfaces with the CSC search system to provide search facilities for use with SDI output and to provide search facilities for user data bases. Use of the cross-over capability, however, is completely optional--SDI users need not direct their output to PLS files and PLS users need not ever search their libraries. The system consists of several components which are listed below.

- PLSXT--a program which collects output from SDI profiles, reformats it, and moves it into the Interface Library
- PRILIB--a collection of program modules to create, expand, maintain, condense, and list libraries of citations
- Conversion Programs--a set of programs used to build libraries from existing machine-readable data bases
- PLSST--a program to reformat a PLS file for searching
- Interface Library--a PLS-format library of citations collected from the search system (both SDI and PLS searches) but not yet merged into User Libraries
- User Libraries--a set of PLS-format libraries associated with individual users, where one library may hold citations for many users or one user may own several libraries

A library is an OS data set containing citations in PLS format. This format is derived from IITRI standard format by adding the user ID number to both directory and string portions of the record along with additional internal items:

	Position	Contents
Record 1	1 - 10	User ID number
	11 - 21	Citation number
	22	'1'
	23 - 262	The directory (60 fullword binary numbers)
	263 - 264	Purge date
Record 2	1 - 10	User ID number
	11 - 21	Citation number
	22	'2'
	23 - 4000	String portion of citation

Both records are OS variable-length records, and only the meaningful portion of the second record is actually present. The records in a library are in order according to the first 22 characters of the records.

5.8.2.1 PLSXT

The PLS Extractor Program (PLSXT) is the search-system-to-PLS interface. It reads the file of profile headers created by INPUTR and builds a list of those containing the code "PRI" in the Security field. It then scans the hit and citations files and extracts citations found by the selected profiles. The citations are converted to PLS format, sorted in ascending order on their first 22 characters, and merged (by the same ordering field) into the interface library. This program is reasonably fast using about 20 seconds of CPU time to extract and reformat 200 citations and merge them into a 10,000-citation library.

Since the regular CSC search system is used for searching PLS libraries, this program is responsible for collecting the results of searches of PLS libraries and making them available to PLS for examination and/or storage.

PLSXT in no way interferes with the SDI search system.

5.8.2.2 PLSST

The PLS Search Transformation Program (PLSST) is the PLS-to-search-system interface. It reads a PLS-format library, sorts it on characters 11-22 (ignoring the user ID number), and then removes the extraneous information added for PLS use. The resulting file can be searched by the standard CSC search system. The results of the search can be re-entered into PLS by coding PRI in the profile(s) used for the search. This program runs rather faster than PLSXT, since only the format conversion and sorting is done.

5.8.2.3 Special Conversion Programs

While PRILIB (below) supports addition of citations to libraries, large collections of data can be added more efficiently by independent conversion programs. One of these programs is for use with a standard card format, and is used for entering data bases which are not in machine-readable form. Other users can be accomodated by special programs written to suit their specific data base, using a combiner/writer module common to all conversion programs. Use of this latter routine insures consistent output.

Running times for these programs varies with the complexity of the format being read. The range of speeds is similar to that for the CSC FORCON's, though they tend to be somewhat faster because sorting the data base involved is usually simpler than sorting a commercial data base. (However, this may not always be true if the data base requires unusual and difficult conversion).

5.8.2.4 PRILIB

The PLS core system contains a group of program modules. The command interface module reads and passes user commands and calls the various modules necessary to perform the desired action. The modules include:

- Command Interface
- Input--performs LOAD and MERGE operations

- Maintenance--performs add, delete, and alter operations
- LISTMON--generates output for one citation
- LISTOUT--combines the results of all LISTMON calls into a single listing
- ACCT--keeps track of user statistics

In use, a library file is read into a temporary disk file that becomes the current file. All manipulation is done with the current file. Other libraries can be merged into it and it can be written, in whole or in part, to create new libraries or replace old ones. Maintenance operations can be performed on citations or on groups of citations and various listings can be generated from all or part of the current file.

5.8.3 User Interaction with Files--Commands

User commands are file commands (for input, output, and listing generation) and maintenance commands (for adding, altering, or deleting citations or fields within citations). File commands are free-format, consisting of an operation keyword, a file name, an ID mask, and a listing command. All of these except the keyword are optional. The ID mask specifies portions of the ID number which must be matched for a record to be read or written. The listing command specifies how the records selected by the ID mask are to be listed, if at all. Listing types are bibliographic, tabular, and keyword-in- and out-of-context (KWIC and KWOC); various sort options are available. Commands available are LOAD and MERGE--for input, and PURGE, DUMP, and EXTRACT--for output. The output commands differ in their effect on the current file--PURGE retains all except selected records, EXTRACT retains only selected records, and DUMP retains the previous current file intact. All three write selected records to the specified file.

Maintenance commands are fixed-format commands consisting of a keyword, ID and citation number mask, a field specifier, and a new data value (except for deletes). If complete ID and citation numbers are given, the program scans forward to the specified citation, performs the specified operations, and leaves the current file where it is, so that if the new commands call for a later citation, previous ones need not be re-scanned. If a command contains "don't care" positions, specified by asterisks, the current file is reset to the beginning and all records matching the required portion are modified or deleted as specified (obviously, "don't care" positions are not permitted in commands to add citations, through they are permitted in commands to add fields). The command formulas are:

<u>File Commands</u>

```
<command>       % ⟨keyword⟩b̸ ⟨specifier⟩
<keyword>       LOAD | MERGE | DUMP | EXTRACT
                PURGE | ABSET | MAINT
<specifier>     ⟨file spec⟩ ⟨mask spec⟩⟨list spec⟩
<file spec>     ⟨file name⟩ |Φ
<mask spec>     , ⟨ID mask⟩ |Φ
<list spec>     / ⟨list type⟩  ⟨sort option⟩
<list type>     TAB | BIBLIO | KWIC | KWOC | USER
<sort option>   ( ⟨field type⟩ | Φ  )
<field type>    AUTHOR | TITLE | CITNO | IDNO
                CODEN | SUBJECT | ⟨type no⟩
<ID mask>       string of ten. Or fewer positions
<file name>     legal OS/360 ddname
<type no>       number in the range 0-999
```

<u>Examples</u>'

```
% LOAD          INFILE,**A01/TAB  (AUTHOR)
% PURGE         OUTFILE,**AO*
% MERGE         UPDFILE/BIBLIO
```

Maintenance commands

Positions	Contents
1 - 10	ID number
11 - 21	Citation number
22 - 25	Keyword
26 - 28	Field type
29 - 30	Iteration of field
31 - 78	Data
79 - 80	Sequence number (for continuations)

Both file and maintenance commands may be continued. For file commands continuation cards begin %b̸, and position 3 of card n+1 is treated as following position 80 of card n. For maintenance commands positions 1-29 of continuation cards match the first card, and positions 79-80 contain ascending numbers.

Two special commands are written in file command format, but specify conditional, rather than immediate output. ABSET is a DUMP command to be performed if PRILIB abends or if a command error causes the job to stop before all commands are fulfilled. MAINT is a DUMP command performed as maintenance operations occur, allowing saving of pre-modification values and generation of listings during maintenance operations.

5. 8 .4 Libraries

PLS libraries are OS sequential data sets. Currently they have record length 4004, blocksize 4008, and variable-blocked format, but shorter lengths could be used. Typical citations obtained from SDI runs on CA Condensates total about 600 bytes, from EI Compendex around 1000 bytes. Frequently-used files might be kept on disk, but most will be tape-resident. CA Condensates citations would fit about 12 per track (on a 2314) if a different blocksize were used to permit more blocks per track. The interface library, in particular, might be disk-resident. In normal use PLSXT merges citations into this file after each search run;

the citations are deleted from the file as users extract them for their personal libraries.

While most libraries consist of bibliographic citations (hence the term libraries) there is no restriction on the contents of citations. Any character-string data can be stored, including numerical values. Listings of citations including numerical data can be generated as with other data, though no arithmetic or totalling operations are supported. Users are permitted to add their own data types as desired, subject to CSC conventions, and define special listing formats to be generated by using the USER listing type.

5.8.5 Use of the System

The Private Libraries System affords users a unique ability to store their SDI output in machine-readable form and access it in various ways, to mix SDI output with other data, to add data to SDI citations, and to search collections of SDI and other citations with CSC profiles. As well as permitting grouped listings, as opposed to individual cards, this permits additional use of the data. In many cases the utility of the data can be increased by adding further information to citations. If, for instance, a file is used as a library catalog, such data as accession numbers, shelf locations, periodical renewal dates, and reader comments might be added to citations. An example of an added field is the secondary citation field which PLSXT adds to citations it creates to refer to the journal issue in which the citation was found.

PLS strikes a careful balance between flexibility and simplicity. While the file and listing commands support the basic data base operations, stand-alone conversion programs, the USER listing type, and user-added data fields permit sufficient flexibility to meet a wide variety of data base needs. A serendipitous side-effect of the modular design and simple data structure is that the program can easily be modified to suit special data base needs. Many common features that would add needless complexity can be built in for special applications. This combination of flexibility and simplicity provides maximum ability at minimum cost.

6. DATA BASES--CHARACTERISTICS, STATISTICS, AND COMPARISONS

In addition to the obvious intended variation in content, data bases vary within external characteristics both within and between supplier organizations. The variation exists in terms of machine code, character code, tape density, labeling conventions, blocking factors, content of logical and physical records, data elements included, data element content, codes employed and format of the tape. CSC analyzed a number of data bases for these items and presented findings in a paper entitled "Comparison of Document Data Bases" which appeared in the Journal of the American Society for Information Science, Volume 22; No. 5, September-October 1971.[4] Such inconsistencies and non-standard representations are accommodated in the CSC system by use of format conversion preprocessor programs as discussed in Section 5.2.

We have done further analysis of the CSC production data bases CA, BA and EI. We have developed statistics and analyzed them in order to gain insights into the use of the data bases, prepare projections for future storage and searching requirements, etc.

6.1 Data Base Characteristics

6.1.1 Number of Citations per Issue

The number of citations per issue varies from data base to data base and often within a data base. BA Previews, for example, produces a fixed number of citations per issue throughout the volume -- 5835 citations appear on each issue of BA and 7500 on each issue of BIORI.

In the case of CA, over the past three years issues of CA Condensates have contained from 3400+ to 8800+ (see Figures 6-1 through 6-5) with the average in 1969 being 4600+ and the average in 1972 being 6000+. EI issues range from 4400+ to 8300+ citations per month (see Figure 6-6) with a small percentage due to erroneous citations being recycled. (No figure is presented for BA because the number is constant.)

The number of citations directly affects the cost of searching and hence the price we must require for subscriptions. Thus, a 30% increase that has occurred in CA from 1969 to 1972 should imply an increase in subscription fee.

The number of citations affects cost and the cost per citation per issue is relatively constant for a fixed number of profiles. As the number of profiles increases cost/citation/issue will increase because individual citations are evaluated for more profiles.

6.1.2 Statistics on Length of Citations, Data Fields per Citation and Key Words per Citation

Statistics are given for CA in Table 6-7 showing: the number (or average number) of citations on a tape(s) together with the mean, standard deviation, and maximum length of the citations; average number of data fields per citation; average number of key words per citation -- mean, standard deviation, and maximum. Note that the mean length of citations (number of characters) and the number of data fields/citation are increased after CAS added two new data fields -- cross references and patent priority codes. Also, the number of key words/citation increased in the later issues. This is due in part to our inclusion of cross references in the key word portion of the IITRI-formatted tape (because cross references provide subject type information) but it also represents an increase in the number of key words assigned by CAS. Such data base additions affect the center both positively and negatively. They increase the cost of processing a tape but also increase retrieval capability by providing more locators.

6.1.3 Percent Occurrence of Data Types

There are many different data types or data elements present on various data bases. Even within a given data base that specifies use of certain data types the frequency

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71,72

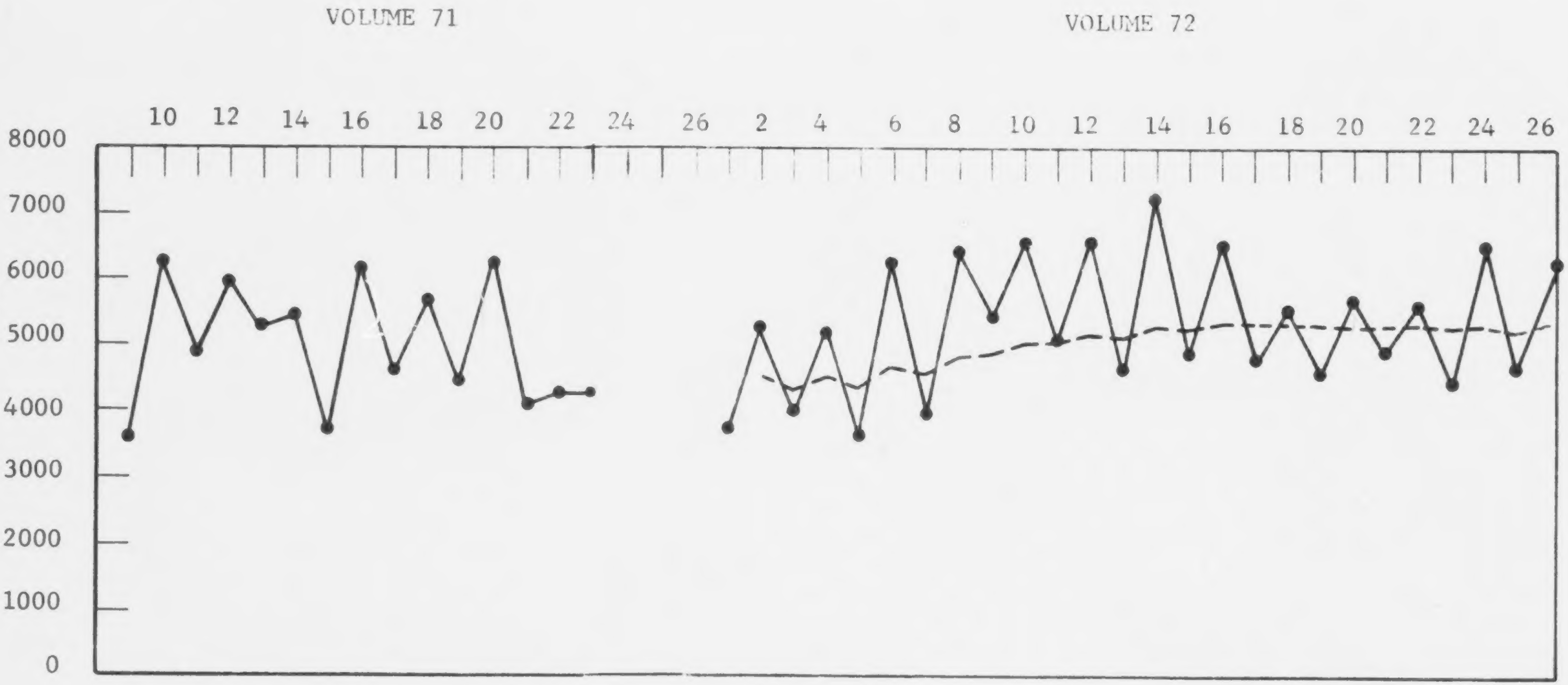


Figure 6-1
NUMBER OF CITATIONS ON TAPE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

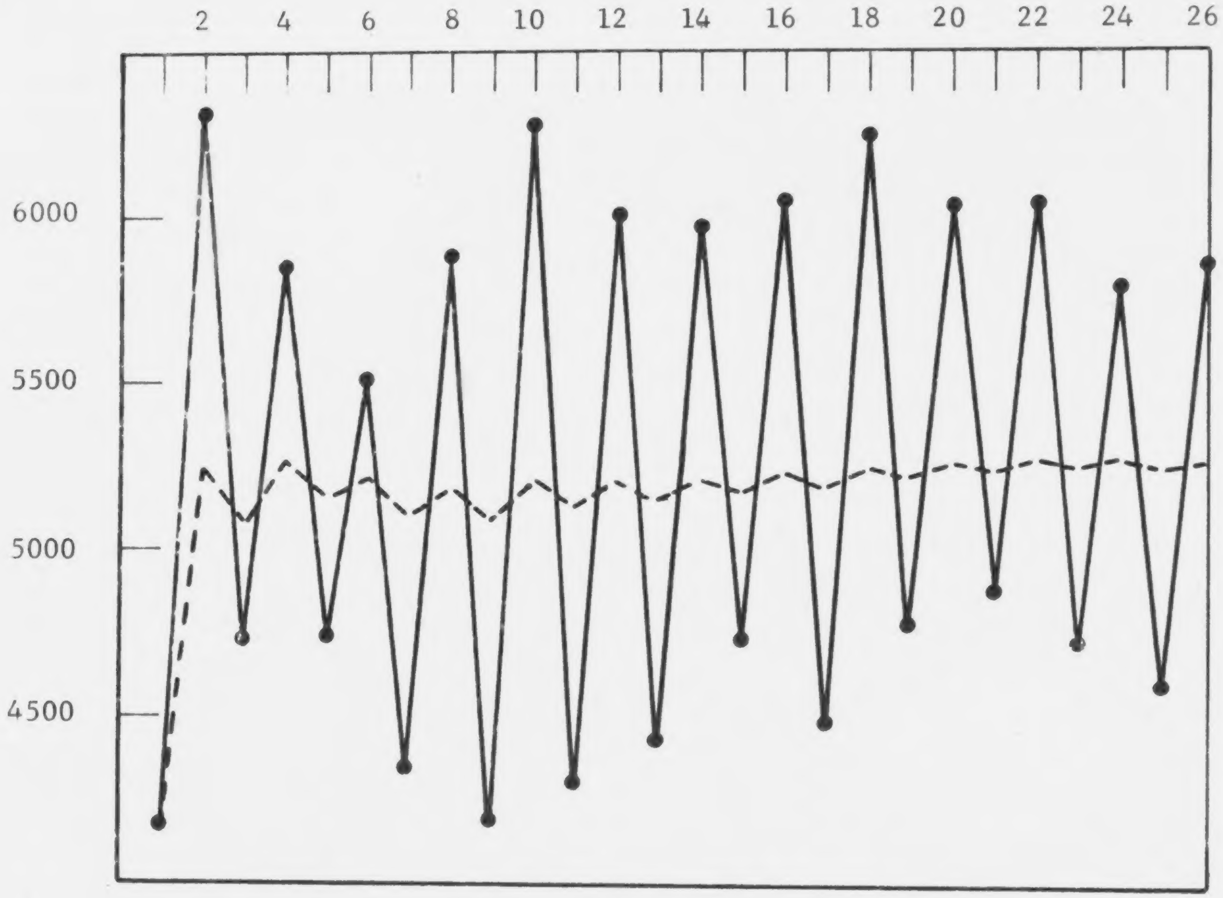


Figure 6-2

NUMBER OF CITATIONS ON TAPE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

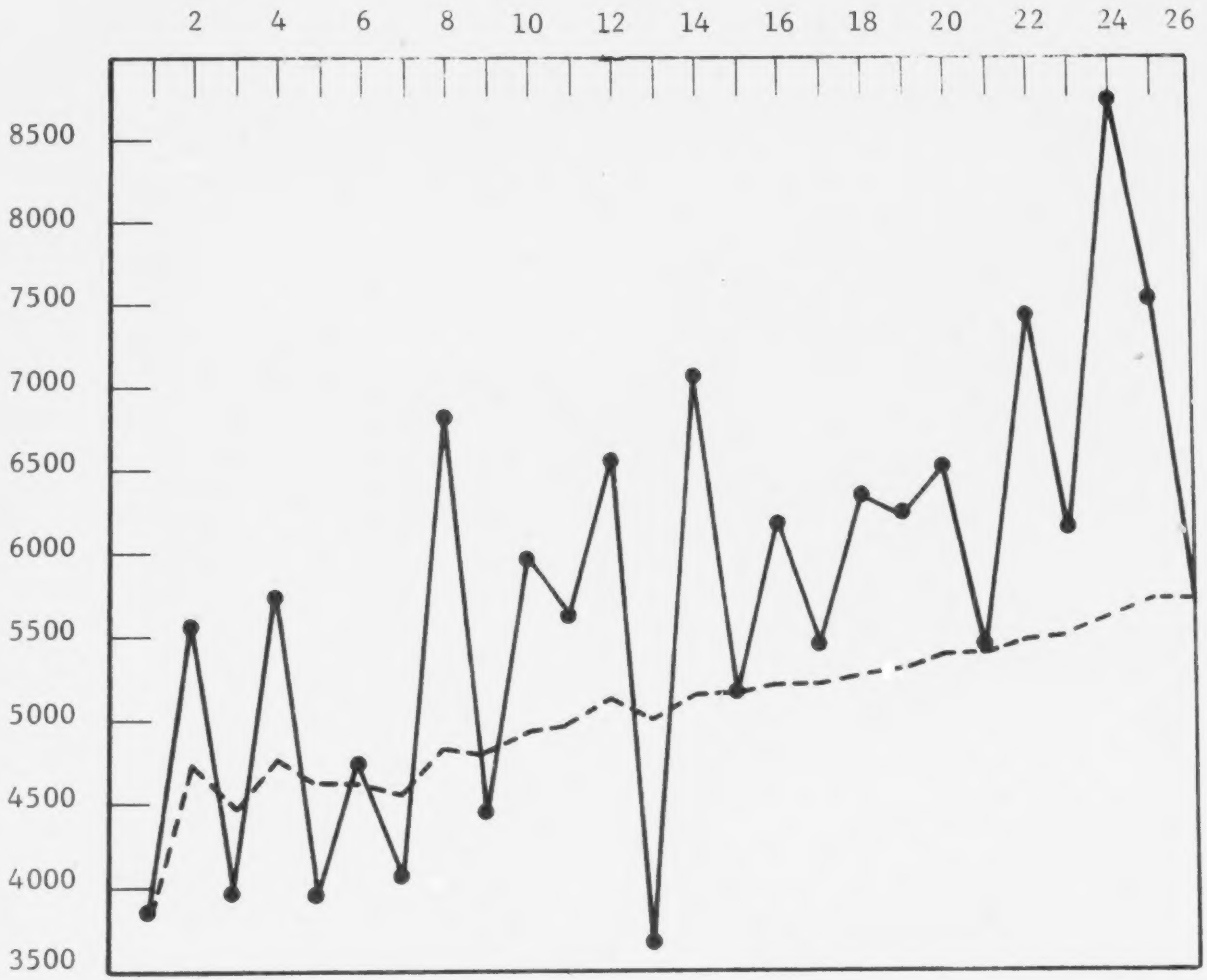


Figure 6-3

NUMBER OF CITATIONS ON TAPE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

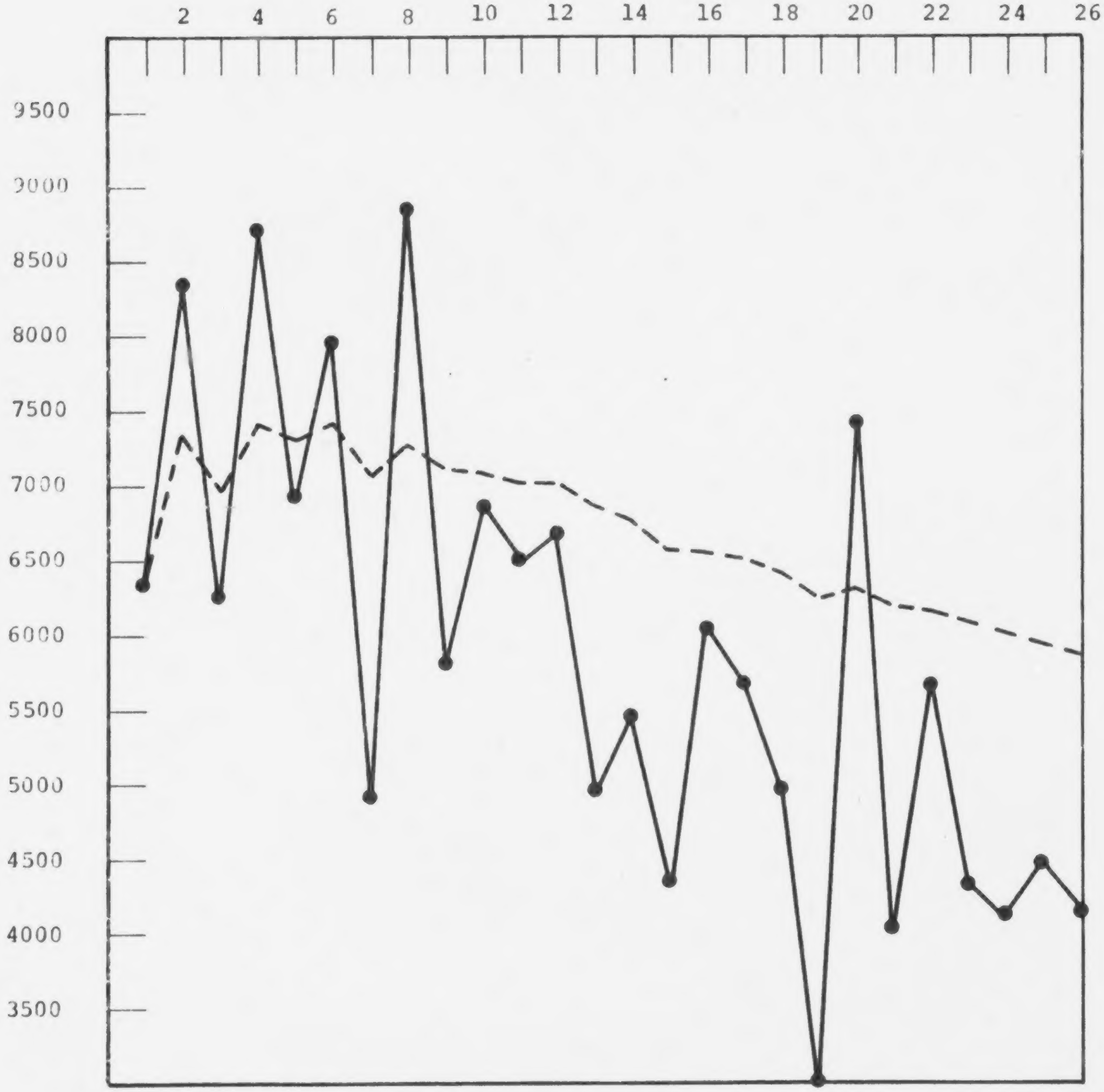


Figure 6-4

NUMBER OF CITATIONS ON TAPE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

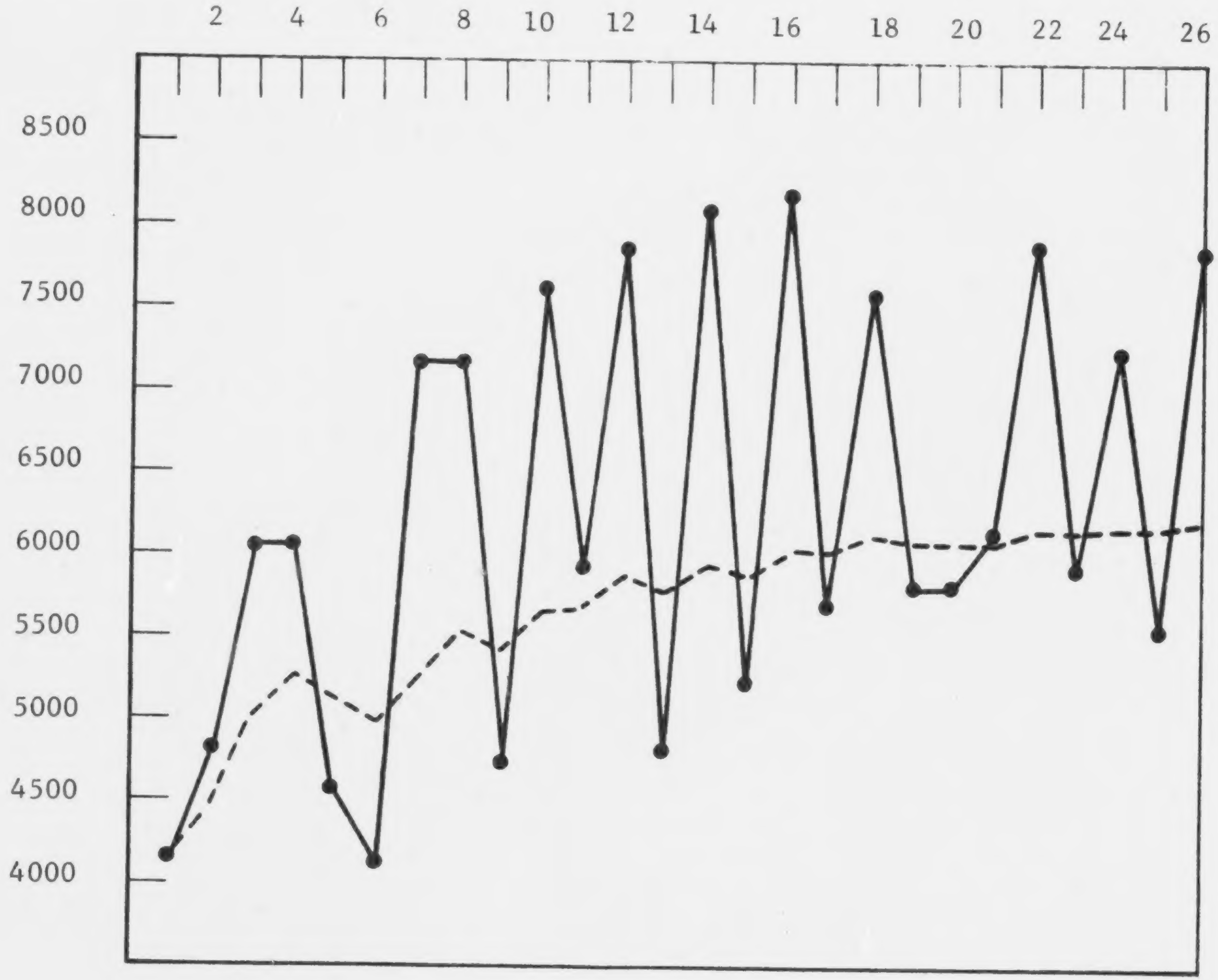


Figure 6-5

NUMBER OF CITATIONS ON TAPE VS. ISSUE

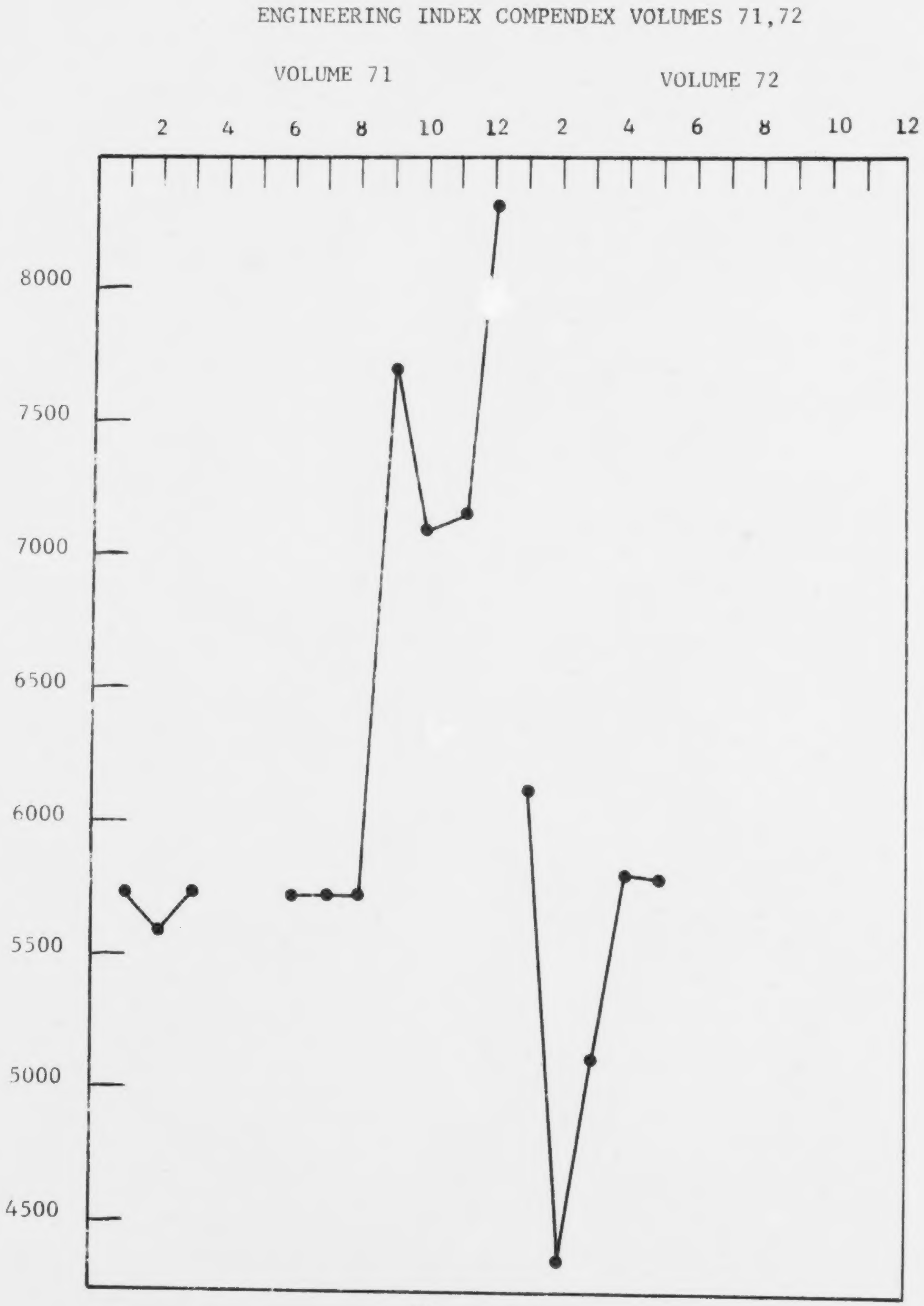


Figure 6-6

NUMBER OF CITATIONS ON TAPE VS. ISSUE

CA CONDENSATES

	Odd-number Issue (CA 76:07*)	Even-number Issue (CA 76:08*)	Odd-number Issues (CA 76:23,25; 77:01)	Even-number Issues (CA 76:24,26)
Citations	4500	7200	5600	7500
Length				
Mean	270	260	280	270
Standard Deviation	50	55	60	58
Maximum	650	600	600	600
Data Fields/-Citation	6.8	6.7	7.1	7.1
Keywords/Citation				
Mean	4.5	5.2	4.9	5.8
Standard Deviation	2	2.5	2.3	3
Maximum	25	35	25	30

* prior to the addition of cross references (type 14) and Patent Priority Codes (type 15) to the data base

Table 6-1

STATISTICS ON LENGTH, DATA FIELDS PER CITATION, AND KEYWORDS PER CITATION IN CA CONDENSATES with which data types appear may vary. In CA, for example, titles, CODEN and keywords are present in 100% of the citations; authors 99%; journal titles 91-98% etc. (See Table 6-2).

6.1.4 Average Length of Data Entries by Data Types

The length (number of characters) of a given data type will vary as can be seen in Table 6-3.

6.2 Data Base Term and Character Occurrences

Phenomena about data bases that affect the way in which profiles should be written are the frequency with which specific terms, letter combinations and letters occur in the data base and the variation between data bases of the occurrences of the same term. In order to monitor growth of vocabulary in data bases, observe differences between data bases and predict the degree of specificity of individual profile terms and truncated terms at the time of writing profiles, CSC has prepared a number of lists for each data base including: term frequency--sorted both alphabetically and in frequency order; KLIC indexes; bigram frequency lists; and single-character frequency lists.[6]

6.2.1 Term Frequency

CSC has developed a program for extracting word tokens from a data base and sorting them both alphabetically and by frequency. These sorted frequency lists are prepared for each data base. Figures 6-7 through 6-12 are samples from CA, BA and EI alphabetical term frequency lists and frequency ordered term frequency lists.

The program, EXTRACT, extracts word tokens from a data base in IITRI format. The type of words (e.g., title words, author words) to be extracted were defined by the programmer. Initially, the program used only blanks. Thus if "X RAY" appears in the data base without a hyphen and with an internal blank, the program will extract "X" and "RAY" as two separate words. Although

CA CONDENSATES

Data Type		Percent Occurrence in			
Number	Name	CA 76:07	CA 76:08	CA 76:23,25	CA 76:24,26
1	CODEN	100	100	100	100
2	Title	100	100	100	100
3	Author	99	99	99	98
4	Journal Title	95	91	98	97
5	Keyword	100	100	100	100
8	Corporate Author	98	96	98	96
13	Availability, etc.	88	83	90	82
14	Cross Reference	0	0	9	16

Table 6-2

PERCENT OCCURRENCE OF DATA TYPES IN
CA CITATIONS

CA CONDENSATES

Data Type		Average Length in			
		CA 76:07	CA 76:08	CA 76:23,25	CA 76:24,26
1	CODEN	28	29	38	30
2	Title	79	72	80	72
3	Author	36	35	37	36
4	Journal Title	19	21	20	20
5	Keyword	41	43	47	50
8	Corporate Author	42	37	43	36
13	Availability, etc.	25	31	24	27
14	Cross Reference	NA	NA	8	8

Table 6-3

AVERAGE LENGTH OF DATA ENTRIES BY DATA TYPE

IN CA CITATIONS

```
ACETYLSTROPHANTHIDIN         1   ACID                      11868
ACETYLSULFAMETHOXYPY         1   ACID.GAMMA                    1
ACETYLSULFANILIC             2   ACID(CHLOROCITRAMALI          1
ACETYLTETRAMETHYLHYD         1   ACID(M)                       1
ACETYLTHEOPHYLLINE           1   ACID(4                        1
ACETYLTHIAZOLIDINE           1   ACID)                         2
ACETYLTHIENYLMERCURY         1   ACID)(TETRAMINE)COBA          1
ACETYLTHIO                   2   ACID>                         4
ACETYLTHIOCHOLINE            1   ACID>)                        1
ACETYLTHIONOCHOLINE          1   ACID:THEOBROMINE              1
ACETYLTHIOPERHYDRO           1   ACIDA                         6
ACETYLTHIOPHENE              3   ACIDAFFIN                     1
ACETYLTRANSFERASE           35   ACIDEMIA                      6
ACETYLTRANSFERASES           1   ACIDEMIAS                     1
ACETYLTROPOLONE              2   ACIDIC                      164
ACETYLTRYPSIN                2   ACIDIFICATION                10
ACETYLTRYPTOPHAN             4   ACIDIFICATON                  1
ACETYLTYROSINE               2   ACIDIFIED                     7
ACETYLUREA                   1   ACIDIFYING                    1
ACETYPICOLINATES             1   ACIDIMETRIC                   6
ACETYSAL                     1   ACIDIMETRY                    3
ACETYSALICYLATE              1   ACIDITIES                    11
ACEYTLENE                    1   ACIDITY                     192
ACHAKSH                      1   ACIDIURICI                    1
ACHAKSK                      1   ACIDIZING                     6
ACHALASIA                    1   ACIDO                         3
ACHATINA                     1   ACIDOCOMPLEX                  1
ACHETA                       5   ACIDOL                        1
ACHIEVE                      2   ACIDOLYSIS                    4
ACHIEVED                     1   ACIDOLYTIC                    1
ACHIEVEMENT                  3   ACIDOPATHIES                  1
ACHIEVEMENTS                 4   ACIDOPENTAAMMINECOBA          2
ACHIEVING                    3   ACIDOPENTAAQUOCHROMI          1
ACHILLEA                     4   ACIDOPENTAMINECOBALT          1
ACHILLENE                    2   ACIDOPHIL                     1
ACHILLENOL                   1   ACIDOPHILA                    1
ACHILLES                     1   ACIDOPHILIC                   5
ACHINSK                      5   ACIDOPHILOUS                  1
ACHIRAL                      1   ACIDOPHILUS                   3
ACHISAI                      1   ACIDOSI                       1
ACHLORHYDRIA                 2   ACIDOSIS                     56
ACHOLEPLASMA                 1   ACIDOTIC                      1
ACHONDRITE                   2   ACIDOTROPHIC                  1
ACHONDRITES                  2   ACIDOVORANS                   1
ACHONDRITIC                  2   ACIDPROOF                     2
ACHONDROPLASTIC              1   ACIDS                      3369
ACHROIA                      1   ACIDS)                        4
ACHROMATIC                   2   ACIDULANT                     2
ACHROMOBACTER               10   ACIDULANTS                    1
ACHROMYCIN                   3   ACIDULATED                    1
ACHTARANDITE                 1   ACIDULATION                   3
ACHYLA                       1   ACIDULENT                     1
ACHYLIA                      1   ACIDULENTS                    1
ACHYRANTHES                  2   ACIDULOUS                     3
ACHYROCLINE                  1   ACIDURIA                      1
```

Figure 6-7

CA ALPHABETICAL TERM FREQUENCY LIST

```
ACETYLPROMAZINE              1
ACETYLTHIO                   1
ACEVALTRATUM                 1
ACHALASIA                    6
ACHAPARRAMIENTO              1
ACHARISTA                    1
ACHATINA                     1
ACHE                         1
ACHEIVEMENTS                 1
ACHELIA                      1
ACHERONTIA                   3
ACHETA                       7
ACHETAL                      1
ACHIEVABLE                   1
ACHIEVED                     3
ACHIEVEMENT                  6
ACHIEVEMENTS                 7
ACHIEVING                    2
ACHILLEA                     3
ACHILLES                     3
ACHILLURBAINIA               2
ACHILLURBAINIIDAE            1
ACHLYA                      13
ACHLYAE                      1
ACHNANTHES                   1
ACHONDROPLASIA               1
ACHONDROPLASTIC              1
ACHRAS                       2
ACHROMATIC                   7
ACHROMATOPSIA                1
ACHROMIANS                   1
ACHROMOBACTER                9
ACHROMYCIN                   1
ACHRUSTERUS                  1
ACHTHERES                    1
ACHYLA                       2
ACHYLIA                      1
ACHYRANTHES                  1
ACICOLA                      2
ACICULA                      1
ACICULARIA                   1
ACICULARIS                   1
ACICULATA                    2
ACID                      1954
ACIDEMIA                     4
ACIDI                        1
ACIDIC                      13
ACIDIFICATION                8
ACIDIFIED                    1
```

Figure 6-8

BA ALPHABETICAL TERM FREQUENCY LIST

ABLENKWINKEL	1	ABSOLUTBESTIMMUNG	1
ABLESUNG	1	ABSOLUTE	68
ABLTATION	1	ABSOLUTELY	2
ABLUFT	1	ABSOLUTESTABILITY	1
ABMAGNETISIERUNGSART	1	ABSORBANCE	2
ABMESSUNGEN	1	ABSORBANT	1
ABNAHME	4	ABSORBANTS:	2
ABNORMAL	9	ABSORBED	12
ABNORMALITIES	2	ABSORBENT	3
ABNORMALLY	1	ABSORBENTS	2
ABOARD	13	ABSORBER	9
ABOBADA	1	ABSORBERN	1
ABORT	1	ABSORBERS	30
ABOSORPTION	1	ABSORBERS:	1
ABOSRPTION	1	ABSORBEUR	1
ABOUT	77	ABSORBIN	1
ABOVE	60	ABSORBING	27
ABOVEGROUND	1	ABSORBS	1
ABPACKEN	1	ABSORP&TION	1
ABPACKMASCHINE	1	ABSORPBERS	1
ABRADED	1	ABSORPTANCE	1
ABRAHAM	2	ABSORPTIOMETER	1
ABRASION	18	ABSORPTIOMETERS	1
ABRASIVE	68	ABSORPTION	640
ABRASIVENESS	1	ABSORPTIONS	1
ABRASIVES	25	ABSORPTIONSSPEKTREN	1
ABRAUMFOERDERBRUECKE	1	ABSORPTION:	2
ABRECHNUNG	1	ABSORPTIVE	4
ABRICHT	1	ABSORPTIVITY	3
ABRIDGE	1	ABSPANEN	1
ABRIDGED	1	ABSPANNTRANSFORMATOR	1
ABRIDGEMENT	1	ABSPERRORGANE	1
ABRIEBFESTER	1	ABSROPTION	2
ABRIEBSBESCHLEUNIGUN	1	ABSTACLE:	1
ABROAD	4	ABSTECHEN	1
ABRONCSSZOVETVAZSK	1	ABSTECKZIEHENS	1
ABRUPT	8	ABSTIMMBAR	1
ABRUPTLY	2	ABSTIMMSAETZE	1
ABS	116	ABSTIMMUNG	1
ABSAUGESCHLITZE	1	ABSTRACT	9
ABSCHAETZUNG	3	ABSTRACTING	7
ABSCHALTUEBERSPANNUN	1	ABSTRACTION	7
ABSCHEIDEGRENZE	1	ABSTRACTS	3
ABSCHEIDELEISTUNG	1	ABSTRAHLUNG	1
ABSCHEIDEN	1	ABTAST	1
ABSCHEIDUNG	3	ABTASTFOLGEWERTEN	1
ABSCHIRMUNG	1	ABTASTGERAET	1
ABSCHMITTSWEISE	1	ABTASTGLIEDER	1
ABSCHNITTSWEISE	2	ABTASTREGELKREISE	2
ABSCHRECKEN	1	ABTASTREGELKREISEN	1
ABSCHRECKGESCHWINDIG	1	ABTASTSYSTEMEN	1
ABSCHRECKHAERTENDER	1	ABTASTVERFAHREN	1
ABSCISSA	1	ABTRAGENDEN	1
ABSENCE	11	ABTRIEBSGESETZ	1

Figure 6-9

EI ALPHABETICAL TERM FREQUENCY LIST

TERM	
HYDROLYSIS	695
DETECTION	694
BARIUM	693
ISOLATED	691
V	688
STRONTIUM	687
ANTIMONY	686
DIFFRACTION	686
IRRADIATION	683
CARBONATE	682
ARSENIDE	680
CONTINUOUS	678
DEPOSIT	677
IODINE	677
BIOCHEMICAL	674
INVESTIGATION	674
BISMUTH	671
CONDUCTIVITY	670
BROMIDE	669
EPOXY	666
ROCK	666
RESINS	664
ARGON	663
ELECTROLYTE	661
PALLADIUM	661
ETHYL	660
DRUGS	659
CARBIDE	654
CHAIN	652
PARAMETERS	652
CATION	650
LATTICE	649
PHENOL	649
CESIUM	648
SOURCE	648
FLUORESCENCE	647
MILK	646
WAVE	644
DYES	643
MODIFIED	643
SPECIES	642
CORRELATION	641
OTHER	638
URINE	637
METHACRYLATE	635
STUDIED	635
GROUPS	633
KIDNEY	633
KINETIC	632
ALKALINE	631
BEAM	631
COATINGS	631

TERM	
FERTILIZER	627
DOPED	626
PLUTONIUM	625
QUANTITATIVE	625
DIELEC	623
PHYSICAL	623
RING	623
QUALITY	621
ADDN	619
CONSTANTS	619
NUTRITION	618
CHANGE	616
CHROMATOGRAPHIC	615
AGE	613
CONFORMATION	613
NITRO	611
EQUIL	610
INSECTICIDE	610
COMPARATIVE	608
IODIDE	608
NATURE	606
SOLNS	606
CONDENSATION	605
CONVERSION	605
NUCLEON	605
FAST	604
SECTION	604
ALPHA	602
TEST	602
FORMALDEHYDE	600
DEPENDENT	599
SILICATE	599
COAL	595
MECHANICAL	595
TRANSFORMATION	595
PROPYLENE	593
HERBICIDE	592
B	591
DETERMINING	591
DEFORMATION	590
PROCESSING	588
RAMAN	588
SPECTRAL	587
MICE	585
YEAST	585
ATOMS	584
IDENTIFICATION	582
SURFACES	582
POWER	579
COLOR	578
COPOLYMERS	578
ELECTRIC	577

Figure 6-10

CA FREQUENCY-ORDERED TERM FREQUENCY LIST

```
                TERM                    #
   GAMMA                               270
   02502                               270
   64010                               270
   MONKEY                              269
   RATE                                269
   TUMORS                              269
   22012                               269
   14502                               268
   16502                               268
   APPLICATION                         267
   CONTRIBUTION                        267
   INSULIN                             267
   MEDICAL                             267
   PHOSPHATASE                         267
   22032                               267
   INDIA                               265
   RECORDS                             264
   62520                               264
   HEMO                                263
   L                                   263
   14002                               263
   14                                  262
   63584                               262
   7                                   262
   PATHOLOGY                           261
   MEAN                                259
   BIOCHEMICAL                         259
   GLOBULIN                            258
   MODEL                               258
   86310                               258
   TRANSFER                            257
   LETTER                              256
   SOUTH                               256
   DIET                                255
   FOREST                              255
   MOLECULAR                           254
   NERVE                               254
   QUALITY                             253
   RELATED                             253
   13018                               253
   CYCLE                               252
   TRI                                 252
   13002                               252
   GASTRIC                             249
   LABORATORY                          249
   VENOUS                              248
   GROUP                               247
   ISOLATION                           247
   TYPE                                247
   26685                               247
   IRRADIATION                         246
   SIGNIFICANCE                        246
```

Figure 6-11

BA FREQUENCY-ORDERED TERM FREQUENCY LIST

TERM		TERM	
USING	1197	3	911
PROCESS	1184	TWO	910
MICROWAVE	1180	MEASURING	907
POLLUTION	1174	DIFFUSION	900
TIME	1168	DISTRIBUTION	895
FIELD	1167	STORAGE	895
LASERS	1159	SOUND	894
WELDING	1154	SCATTERING	893
NICKEL	1142	QUALITY	877
FILM	1138	SPECTRUM	858
ELECTRONIC	1133	MOTORS	854
MACHINERY	1130	SIMULATION	852
LINES	1127	FIELDS	851
GASES	1126	STRUCTURES	848
TUBES	1122	CURRENT	846
INDUSTRY	1121	DEVELOPMENT	845
PHYSICAL	1108	CHEMISTRY	844
USE	1107	INFLUENCE	833
ENERGY	1105	CIRCUIT	832
VIBRATIONS	1101	BEAMS	831
APPLICATION	1098	FATIGUE	819
INFORMATION	1096	INVESTIGATION	818
FREQUENCY	1092	COMMUNICATION	807
PLASMAS	1081	WAVE	807
SILICON	1080	NONLINEAR	802
ENGINES	1079	REINFORCED	799
DETERMINATION	1076	STUDIES	797
X	1069	SOLAR	787
INTEGRATED	1065	METALLURGY	784
FILTERS	1048	ION	782
PROPAGATION	1047	PLASMA	781
OIL	1046	BOUNDARY	775
GLASS	1041	SINGLE	769
RAY	1041	COATINGS	766
LINEAR	1035	SATELLITES	766
CONSTRUCTION	1032	PERFORMANCE	762
VACUUM	1030	AMPLIFIERS	761
PRODUCTION	1024	AERODYNAMICS	760
IRRADIATION	1005	DIE	755
PROBLEMS	990	MODEL	754
LIQUID	988	INFRARED	748
PRODUCTS	972	FRACTURE	745
AS	959	COAL	744
PROCESSES	958	LARGE	744
MACHINES	955	MOLDING	744
STRESS	943	STRENGTH	742
POLYMERIZATION	940	ELECTRONS	737
OPTIMIZATION	939	SOLUTIONS	737
CARBON	935	MATHEMATICS	736
TELEPHONE	933	STEAM	735
BIOENGINEERING	932	COMBUSTION	733
MATHEMATICAL	932	OXIDATION	732

Figure 6-12

EI FREQUENCY-ORDERED TERM FREQUENCY LIST the arbitrary choice of blanks as delimiters resulted in some terms splitting, it appeared to be a realistic convention, since none of the data bases are produced under an explicit set of delimiting conventions that could be incorporated in EXTRACT.

This delimiter was used for samples of 2, 6, and 13 issues of CA Volume 72. However, later analysis showed that a small number of discrete terms could be identified by adding slashes and asterisks as delimiters and this was done for 13 issues of Volume 73. When we prepared the frequency lists for Volume 75 we stripped off non-alphanumeric characters from the beginning and/or ends of words.

The second program, SQUEEZ, compresses the extracted word tokens into a list of word types (unique words), and maintains with each word type a count of the number of times that type was found. SQUEEZ makes use of the IBM SORT/MERGE utility program to sort all extracted words alphabetically. It compares each word to the preceding one in the sorted list, and removes duplicates, counting each time it does so. The alphabetical list is printed out, with the count for each word. (See Figures 6-7, 6-8, and 6-9).

At this point a program called CLEAN strips off non-alphanumeric initial and terminal characters in order to avoid listing such terms as HEAT. and HEAT as separate words.

The program, FREQDT, is used to print out the unique words in decreasing order of their frequency of appearance. Again, the SORT/MERGE program is used to sort by frequency count, and the words are then printed in a one, two, or four column format. (See Figures 6-10, 6-11, and 6-12).

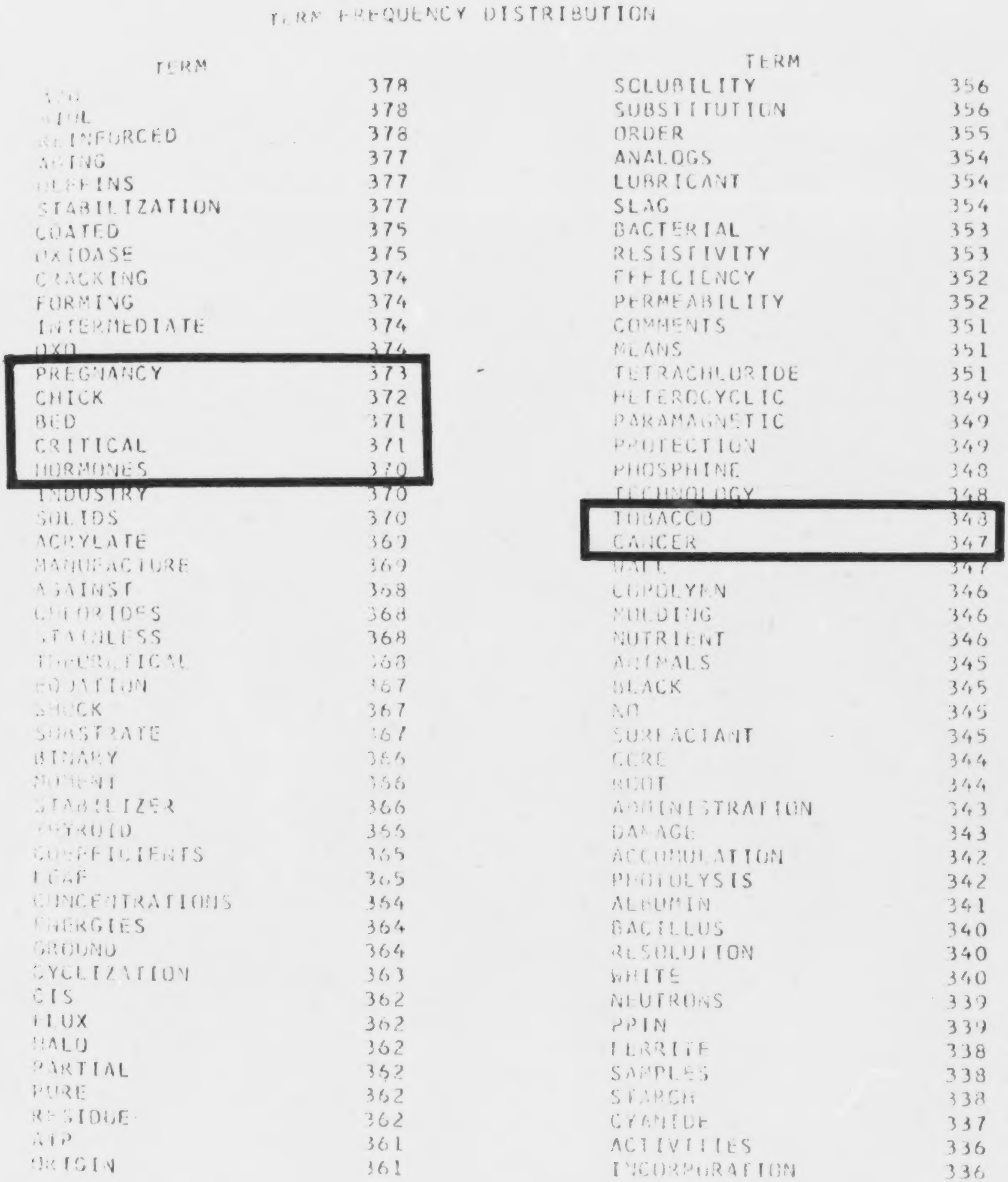
The frequency lists are useful in determining which terms are likely to be highly discriminating because of low frequency and which terms are likely to have poor retrieval effectiveness because of their high frequency. The term frequency lists are intended both as user aids and analytical tools. Occassionally term sequences occur that seem to convey other information, for example, on one page of the list (see Figure 6-13) the terms "tobacco" and "cancer" appear in sequence having frequencies of 348 and 347. On the same page the terms "pregnancy", "chick", "bed", "critical", and "hormones" are listed sequentially having frequencies of 373, 372, 371, 371, and 370.

As might be expected, the prepositions and conjunctions are of high frequency, but within the twenty-five words of highest frequency are also: EFFECT, REVIEW, ACID, ACIDS, DETERMINATION, CHEMICAL, STRUCTURE, PROPERTIES, IRON, and SYNTHESIS. Some of these terms can be used for search terms, but should be used with care, since they could result in hits on a large portion of the file. They should be qualified by incorporation in phrases or associated with other terms in the logic expression.

6.2.2 Type:Token Ratios

After preparing frequency data we analyzed them to determine the number of occurrences (tokens) of unique terms (types). For CA, we did a series of these studies, using 2, 6, and 13 issues of Volume 72 and 13 issues of Volume 73 making a total of 26 issues. In this way we could get a curve of type:token ratio versus tokens. As would be expected, the type:token ratio increases with an increased number of citations. Each type appears, on the average, 5.48 times in 9000 citations taken from two issues, but 12 times for 134,000 citations taken from 26 issues. A summary is given in Table 6-4. The curve of type:token ratio versus tokens, plotted on a log scale, is a straight line (see Figure 6-14).

Although it is probably not reasonable to project this line, if such is done the indications are that no new types would be added once the data base reached 45 million tokens (about 12 years worth of CA) and we know that there are approximately 100,000 new compounds (which have names that may be reported in the literature) developed each year and there are likely to be newly coined words in a growing and changing technological society.



```
                        TERM FREQUENCY DISTRIBUTION

          TERM                                          TERM
[illegible]                378           SOLUBILITY                 356
[illegible]                378           SUBSTITUTION               356
REINFORCED                 378           ORDER                      355
AGING                      377           ANALOGS                    354
OLEFINS                    377           LUBRICANT                  354
STABILIZATION              377           SLAG                       354
COATED                     375           BACTERIAL                  353
OXIDASE                    375           RESISTIVITY                353
CRACKING                   374           EFFICIENCY                 352
FORMING                    374           PERMEABILITY               352
INTERMEDIATE               374           COMMENTS                   351
OXO                        374           MEANS                      351
PREGNANCY                  373           TETRACHLORIDE              351
CHICK                      372           HETEROCYCLIC               349
BED                        371           PARAMAGNETIC               349
CRITICAL                   371           PROTECTION                 349
HORMONES                   370           PHOSPHINE                  348
INDUSTRY                   370           TECHNOLOGY                 348
SOLIDS                     370           TOBACCO                    348
ACRYLATE                   369           CANCER                     347
MANUFACTURE                369           WALL                       347
AGAINST                    368           COPOLYMN                   346
CHLORIDES                  368           MOLDING                    346
STAINLESS                  368           NUTRIENT                   346
THEORETICAL                368           ANIMALS                    345
ROTATION                   367           BLACK                      345
SHOCK                      367           NO                         345
SUBSTRATE                  367           SURFACTANT                 345
BINARY                     366           CORE                       344
MOMENT                     366           ROOT                       344
STABILIZER                 366           ADMINISTRATION             343
THYROID                    365           DAMAGE                     343
COEFFICIENTS               365           ACCUMULATION               342
LEAF                       365           PHOTOLYSIS                 342
CONCENTRATIONS             364           ALBUMIN                    341
ENERGIES                   364           BACILLUS                   340
GROUND                     364           RESOLUTION                 340
CYCLIZATION                363           WHITE                      340
CIS                        362           NEUTRONS                   339
FLUX                       362           PPIN                       339
HALO                       362           FERRITE                    338
PARTIAL                    362           SAMPLES                    338
PURE                       362           STARCH                     338
RESIDUE                    362           CYANIDE                    337
ATP                        361           ACTIVITIES                 336
ORIGIN                     361           INCORPORATION              336
```

Figure 6-13

TERM SETS WITH SIMILAR FREQUENCIES

Number of Issues (in CA Vol.)	Citations	Tokens	Types	Type/Token Ratio
2(Vol. 72)	9,067	91,760	16,753	1:5.48
6(Vol. 72)	31,402	479,856	60,876	1:7.88
13(Vol. 72)	67,456	877,734	92,216	1:9.52
13(Vol. 73)	66,796	963,698	100,498	1:9.59
26(Vol. 72 and 73)	134,252	1,841,432	153,268	1:12.01

Table 6-4
CA TYPE:TOKEN RELATIONSHIPS

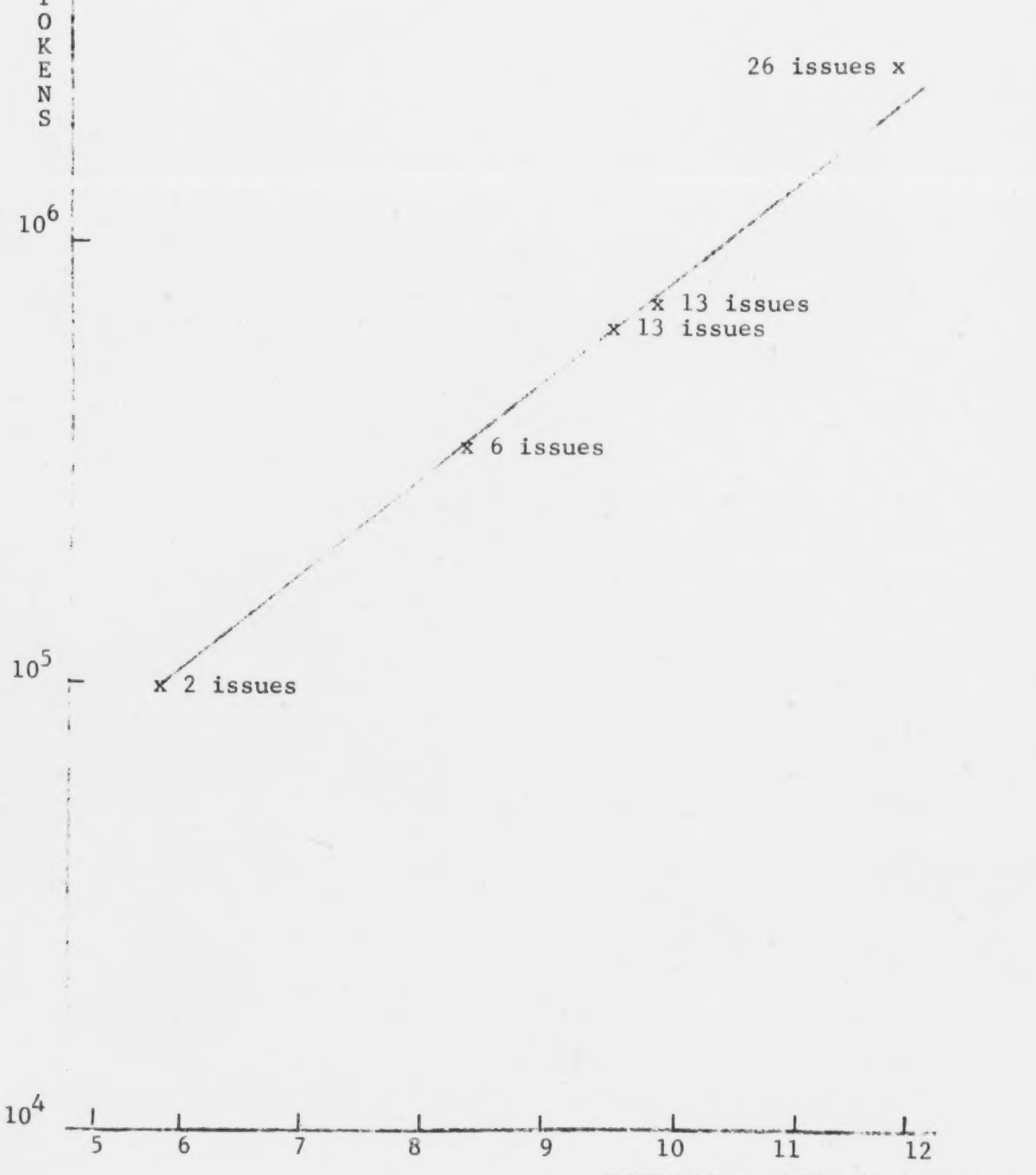


Figure 6-14

REDUCTION IN TYPE:TOKEN RATIO AS A FUNCTION OF NUMBER OF TOKENS

Although the absolute number of term types does not increase linearly, the rate of increase is constant.

The type:token ratios for the earlier volumes of CA (72 and 73) vs. Volume 75 differ probably due to our stripping off non-alphanumeric characters via the CLEAN program when we ran Volume 75. (See Table 6-5).

CA Volumes	No. Issues	No. Types	No. Tokens	Type:Token Ratio
72 & 73	26	153,268	1,841,432	1:12.01
75	26	100,220	2,217,158	1:22.12

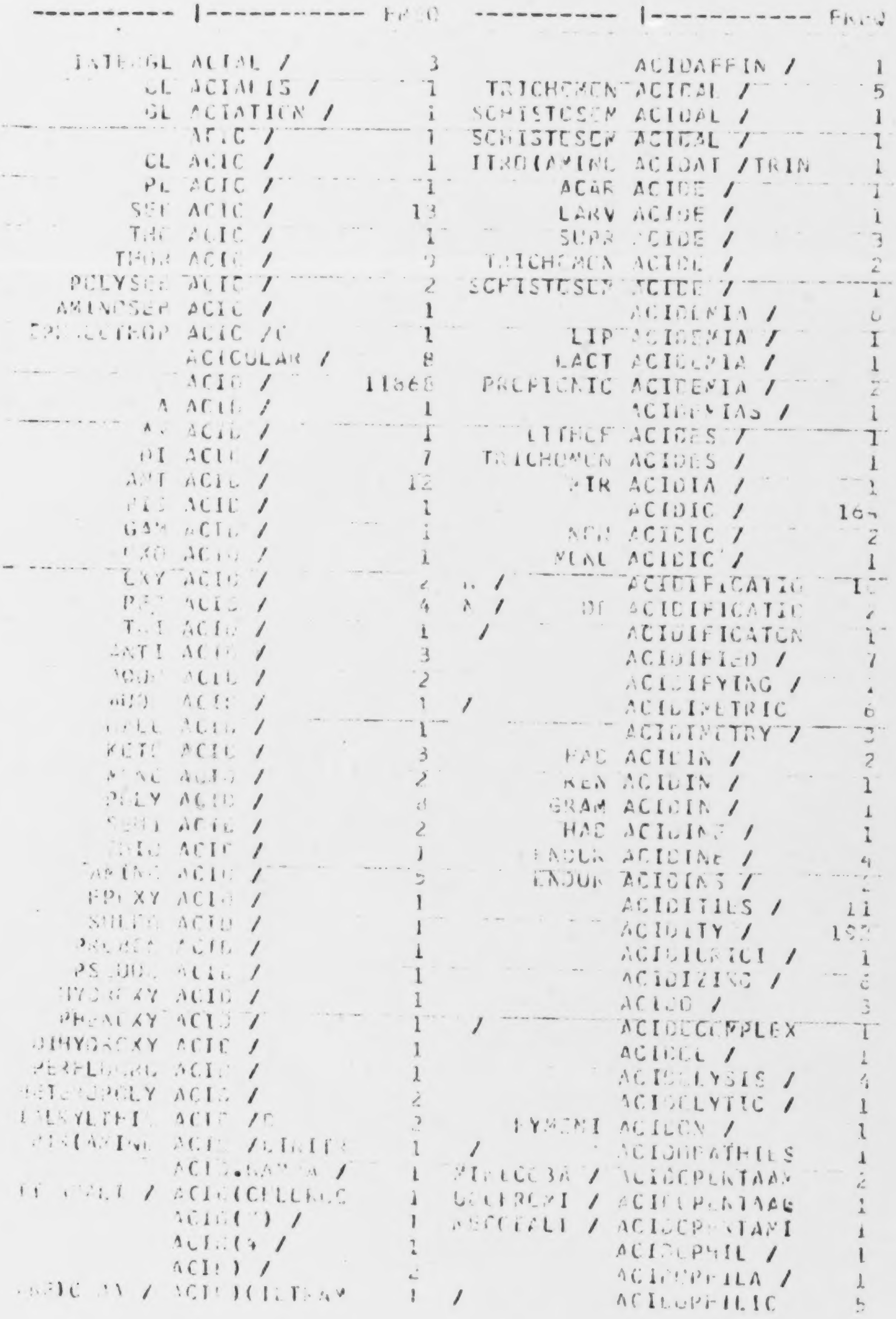
Table 6-5

TYPE:TOKEN RATIOS

6.2.3 Key-Letter-in-Context Listings

The data base analysis programs discussed above; EXTRACT, SQUEEZ, CLEAN, and FREQDT; are followed by a fifth program, KLICPT, which generates a Key-Letter-In-Context (KLIC) index. A KLIC is a permuted word listing sorting on each letter in each word in the data base with the remainder of the word wrapped around it (similar to a KWIC index). The KLIC index is printed with the term frequency following each term. Figures 6-15, 6-16, and 6-17 are sample pages from CA, BA, and EI KLICs.

The KLIC for CA Volume 75 contained 26 issues from July 1, 1971 through December 31, 1971 and contained 157,995 citations. The EXTRACT program extracted 2,217,158 words from the title and keyword fields. 73,470 contained non-alphanumeric initial and terminal characters that were stripped off by the CLEAN program. The SQUEEZ program selected 100,220 unique words. The 1,097,512 KLIC Index entries were sorted in the 12th position of a 20 position field and KLICPT was run to write the KLIC for printing.



```
---------- |----------- FREQ   ---------- |----------- FREQ

    INTEGGL ACIAL /          3              ACIDAFFIN /      1
         GL ACIALIS /        1     TRICHOMON ACIDAL /        5
         GL ACIATION /       1    SCHISTOSOM ACIDAL /        1
            ACIC /           1    SCHISTOSOM ACIDAL /        1
         GL ACIC /           1    ITRO(AMINO ACIDAT /TRIN    1
         PL ACIC /           1          ACAR ACIDE /         1
        SEC ACIC /          13          LARV ACIDE /         1
        THE ACIC /           1          SUPR ACIDE /         3
       THOR ACIC /           9     TRICHOMON ACIDE /         2
    POLYSOR ACIC /           2    SCHISTOSOM ACIDE /         1
   AMINOSER ACIC /           1               ACIDEMIA /      6
[illegible]ECTROP ACIC /C    1           LIP ACIDEMIA /      1
            ACICULAR /       8          LACT ACIDEMIA /      1
            ACID /       11868     PROPIONIC ACIDEMIA /      2
          A ACID /           1               ACIDEMIAS /     1
         A? ACID /           1        LITHOF ACIDES /        1
         DI ACID /           7     TRICHOMON ACIDES /        1
        ANT ACID /          12           MIR ACIDIA /        1
        FI? ACID /           1               ACIDIC /      16?
        GAM ACID /           1           NON ACIDIC /        2
        EXO ACID /           1          MEAL ACIDIC /        1
        OXY ACID /           2 N /           ACIDIFICATIO   1?
        PER ACID /           4 N /        DE ACIDIFICATIO    2
        TRI ACID /           1  /            ACIDIFICATON    1
       ANTI ACID /           3               ACIDIFIED /     7
        NOR ACID /           2               ACIDIFYING /    1
        AMO ACID /           1 /             ACIDIMETRIC     6
       GALL ACID /           1               ACIDIMETRY /    3
       KETO ACID /           3           FAC ACIDIN /        2
       AMINO ACID /          2           KEN ACIDIN /        1
       POLY ACID /           8          GRAM ACIDIN /        1
       SEMI ACID /           2           HAD ACIDINE /       1
       DIIO ACID /           1         ENDUR ACIDINE /       4
     AMINO ACID /            5         ENDUR ACIDINS /       2
      EPOXY ACID /           1               ACIDITIES /    11
      SULFO ACID /           1               ACIDITY /     19?
     PROPEN ACID /           1               ACIDIURICI /    1
     PSEUDO ACID /           1               ACIDIZING /     6
    HYDROXY ACID /           1               ACIDO /         3
    PHENOXY ACID /           1  /            ACIDOCOMPLEX    1
  DIHYDROXY ACID /           1               ACIDOL /        1
  PERFLUORO ACID /           1               ACIDOLYSIS /    4
 HETEROPOLY ACID /           2               ACIDOLYTIC /    1
 ALKYLETHE? ACID /P          2        FYMENI ACIDON /        1
  BIS(AMINO ACID /[illegible] 1 /            ACIDOPATHIES    1
            ACID.GAMMA /     1 PIPECOBA / ACIDOPENTAAM       2
[illegible] / ACID(CELLULO   1 DECAPROMI / ACIDOPENTAAC      1
            ACID(*) /        1 ABCOBALT / ACIDOPENTAMI       1
            ACID(+ /         1               ACIDOPHIL /     1
            ACID) /          2               ACIDOPHILA /    1
[illegible]IC AN / ACID)CITRAM 1 /           ACIDOPHILIC     5
```

Figure 6-15

CA KEY-LETTER-IN-CONTEXT INDEX

```
                 KEY-LETTER-IN-CONTEXT INDEX

  ---------- |----------        ---------- |----------

          TR ACHYTIONE /                    ACID /
           P ACHYTREMA /               LEYR ACID /
      POLYST ACHYUM /                  SUPP ACID /
    PLECTOST ACHYUM /                 AMINO ACID /
           T ACHYURA /                EUMAST ACID /
          BR ACHYURA /                   PL ACIDA /
          BR ACHYURAN /                 AST ACIDAE /
          BR ACHYURUS /                 LIM ACIDAE /
    PLECTOST ACHYUS /                 DELPH ACIDAE /

       MEMBR ACIDAE /                     P ACIFIC /
      EUMAST ACIDAE /                     P ACIFICA /
      GONEPL ACIDAE /           /        OP ACIFICATION
      SCOLOP ACIDAE /                 INDOP ACIFICULA /
  HALACROCOR ACIDAE /P                    P ACIFICUM /
    TRICOMON ACIDAL /                     P ACIFICUS /
   TRICHOMON ACIDAL /                    OP ACIFIED /
        SUPR ACIDE /                   HIER ACIFOLIA /
      TERMIT ACIDE /                     AC ACIIN /
             ACIDEMIA /                  UR ACIL /
        LACT ACIDEMIA /                 LEN ACIL /
       AMINO ACIDEMIA /                 MET ACIL /
             ACIDI /                    MIR ACIL /
         MIR ACIDIA /                  BROM ACIL /
             ACIDIC /                  HERP ACIL /
 N /         ACIDIFICATIO              TERB ACIL /
             ACIDIFIED /                    ACILA /
             ACIDIFYING /                GR ACILARIA /
             ACIDITY /                    F ACILE /
         MIR ACIDIUM /                   GR ACILE /
             ACIDO /                      M ACILENTA /
      HYMENI ACIDON /                    GR ACILENTA /
             ACIDOPATHY /              INFR ACILIATURE /
             ACIDOPHILI /     /          GR ACILICORNIS
 /     ACETO ACIDOPHILUM                 GR ACILIMANUS /
 /           ACIDOPHILUS                  F ACILIS /
             ACIDOSES /                  GR ACILIS /
             ACIDOSIS /                   F ACILITATE /
        KETO ACIDOSIS /                   F ACILITATED /
        LACT ACIDOSIS /                   F ACILITATES /
 /           ACIDOVORANS      /           F ACILITATING
```

Figure 6-16

BA KEY-LETTER-IN-CONTEXT INDEX

```
----------- |----------- FREQ   ---------- |----------- FREQ

REICHSBEOB ACHTUNGE /BE      1           D' ACIDE /          1
      BEOB ACHTUNGEN /       3   POLYESTER ACIDE /          1
      BEOH ACHTUNGEN /       1             ACIDENT /        1
      BETR ACHTUNGEN /       9             ACIDES /         1
TUR / BEOB ACHTUNGSAPER      1             ACIDI /          1
/   LUFTFR ACHTVERKEHR       1             ACIDIC /         5
UNGSUEBERW ACHU /STRAHL      1  N /        ACIDIFICATIO     2
IONSUEBERW ACHUN /EMISS      1             ACIDITY /        6
    UEBERW ACHUNG /          5             ACIDIZING /      1
IERSUEBERW ACHUNG /BETR      1             ACIDS /        103
FOERUBBARM ACHUNG /WI        1       ESTER ACIDS /          1
CH /UEBERW ACHUNGSEINRI      1     HYDROXY ACIDS /          1
      MASS ACHUSETTS /       1  AMINOAMIDO ACIDS /POLY      1
       PAR ACHUTE /          5             ACIDS: /         3
       PAR ACHUTES /        13             ACID: /          2
       PAR ACHUTING /       12        OXID ACIE /           1
NG /     N ACHVERDICHTU      1      DEFORM ACIE /           1
/        S ACHVERHALTE       1             ACIENCE /        1
         N ACHWACHSEN /      1             ACIER /          9
         N ACHWAERMEN /      1          D' ACIER /          6
    WARMFL ACHWALZEN /       2          GL ACIER /          4
         N ACHWEIS /         6          L' ACIER /          5
STICKEITSN ACHWEISE /FE      1             ACIERS /        29
CHERHEITSN ACHWEISE /SI      1          D' ACIERS /         8
DEN /    N ACHWEISMETHO      1          GL ACIERS /        12
 /       N ACHWIRKUNGEN      1           M ACIERZ /         1
        BL ACHY /            1           F ACIES /          2
 /       T ACHYSTOSCOPE      1       ACCUR ACIES /          3
S /      T ACHYSTOSCOPE      1      LITHOF ACIES /          1
        TR ACHYTE /          1           F ACIES) /         1
    TASKTE ACH: /            1           P ACIFIC /        49
           ACI /             5           P ACIFIC: /        3
         N ACI /             1          OP ACIFIED /        2
     NUKLE ACI /             2        OXID ACII /           1
    VALCOV ACI /             1         LOK ACIJ /           1
   ZATEZOV ACI /             2       REGUL ACIJ /           1
SULL'EFFIC ACIA /            1       REGUL ACIJA /          2
    PERIGL ACIAIRE /         1  ELEKOMUNIK ACIJAH /T        1
         F ACIAL /           2       SIMUL ACIJO /          1
        GL ACIAL /           3  ELEKOMUNIK ACIJSKI /T       1
     NAVIE ACIAL /           1       IONIZ ACIJSKIH /       1
    INTERF ACIAL /          63     KOMPENZ ACIJSKIH /       1
    INTERF ACIALE /          1       INDIK ACIJU /          1
    INTERF ACIALES /         1          UR ACIL /           2
      LAPL ACIAN /           4        BROM ACIL /           1
        GL ACIATED /         1     VINYLUR ACIL /           1
       BOR ACIC /            1          GR ACILIS /         1
      THOR ACIC /            3           F ACILITATE /      3
           ACICULAR /        5           F ACILITATED /     1
           ACID /          423  /        F ACILITATING      1
         / ACID /            1           F ACILITIES /     79
        DI ACID /            1           F ACILITIES: /     2
```

Figure 6-17

EI KEY-LETTER-IN-CONTEXT INDEX

The EI Frequency Lists and KLIC generated from EI COMPENDEX Volume 71, issues 1, 2, 3, 6, 7, 8, 9, 10, 11, and 12 contained 1,096,994 words taken from titles and index terms. Of these 115,669 were stripped of terminal punctuation. Redundant words were removed yielding 54,914 unique words for a Type:Token ratio of 1:20. The number of KLIC index entries was 515,317. Following the KLIC the EI bigram frequencies were prepared.

6.2.4 Bigram Frequencies

The CSC search system employs a Least Common Bigram (LCB) technique (Section 5.6). LCB's depend on a bigram (2 letter combination) frequency list which is prepared following the KLIC index. An alphabetical list of bigrams (with frequencies) is prepared as the last step in KLICPT. (See Figure 6-18). A small program, PRTLCB (Print LCB's), was written to print out bigrams in 4 column order. One column is printed for each of four bigram files (BA, Volume 73 CA, EI and Volume 75 CA). If SORT/MERGE is run before PRTLCB, the listing is generated in decreasing frequency order. (See Figure 6-19).

Bigram frequency lists are prepared for each data base. When printed in frequency order they can be looked at as LCB lists. Many of the LCB's for CA, BA, and EI rank as low frequency bigrams in each data base but their position in terms of frequency differ a bit as can be seen in Table 6-6 where bigram frequencies for CA Volume 73, EI Volume 71 and BA Volume 52 were compared with CA Volume 75.

6.2.5 Single Character Frequencies

Another small program, CHRCNT (Character Count), is used to generate a listing of single-character frequencies. The normal listing is alphabetical, but again, if SORT/MERGE is used as a prefatory step, the output can be obtained in frequency order.

The frequency of occurrence single characters as they appear in CA, BA, and EI are:

- for BA (based on Vol. 52)
  ␢0EAIOTN1SR52CL6MHDU4PF38GY7BVW9XKZQJ.$(+;)*=' ?:,/-
- for EI (based on Vol. 71)
  EIT␢NARSOLCUMDPGHFYBVWKXZQ0J,192'635748)(;.+= ?:/-*$

NO.	BA 52		CA 73		EI 71		CA 75	
540	RS	5607	CB	25	C:	235	D&	1
541	RT	12108	CC	4977	D	68681	D)	42
542	RU	12072	CD	645	D<	1	D,	94
543	RV	3132	CE	52647	D(	5	D>	102
544	RW	240	CF	122	D)	6	D:	7
545	RX	7	CG	72	D,	3	D'	40
546	RY	10299	CH	67099	D?	15	D=	2
547	RZ	49	CI	42030	D#	1	D"	1
548	R1	1	CJ	7	D'	609	DA	14963
549	R2	1	CK	9631	DE	2	DB	186
550	R5	2	CL	21124	DA	9245	DC	63
551	S	100485	CM	101	DB	324	DD	4026
552	SA	6775	CN	1372	DC	269	DE	115967
553	SB	210	CO	96196	DD	1031	DF	68
554	SC	8703	CP	54	DE	46579	DG	662
555	SD	33	CQ	309	DF	87	DH	1130
556	SE	27754	CR	20252	DG	1008	DI	95492
557	SF	714	CS	9576	DH	527	DJ	163
558	SG	31	CT	93118	DI	40658	DK	20
559	SH	4050	CU	14678	DJ	111	DL	816
560	SI	25795	CV	23	DK	33	DM	2307
561	SJ	23	CW	10	DL	1015	DN	13511
562	SK	1141	CX	53	DM	527	DO	13475
563	SL	1167	CY	17283	DN	275	DP	459
564	SM	4346	CZ	322	DO	3286	DQ	1
565	SN	226	C0	2	DP	131	DR	31504
566	SO	13095	C1	105	DQ	3	DS	17982
567	SP	13333	C2	71	DR	8288	DT	837
568	SQ	272	C3	44	DS	13547	DU	24145
569	SR	600	C4	34	DT	251	DV	217
570	SS	9617	C5	47	DU	17695	DW	240
571	ST	43445	C6	42	DV	327	DX	56
572	SU	10086	C7	16	DW	418	DY	12081
573	SV	124	C8	19	DX	6	DZ	197
574	SW	569	C9	19	DY	6111	D0	5
575	SY	6764	D	117249	DZ	47	D1	42
576	SZ	23	D.	6110	D0	3	D2	77
577	S1	5	D(	102	D1	2	D3	127
578	S3	1	D+	15	D3	2	D4	27
579	S4	1	D$	11	D4	1	D5	16
580	S6	1	D*	1	D9	1	D6	14

Figure 6-18

ALPHABETICAL BIGRAM LISTS

NO.	BA 52		CA 73		EI 71		CA 75	
1	CO	185506	S	333266	S	239242	E	353092
2	1	160537	IN	250635	IN	149186	N	333473
3	50	120594	N	240680	ON	149047	IN	309011
4	E	101814	EN	234477	TI	148834	ON	301243
5	S	100485	E	228943	N	133065	S	301010
6	10	98219	TI	211554	ER	131582	TI	269159
7	N	96583	A	205516	E	129378	AT	216395
8	06	93233	ES	174195	AT	112989	AN	204034
9	IN	85818	AT	170045	TE	102934	ER	191437
10	4	80470	AN	168902	IC	102405	RO	191175
11	C	80203	C	166145	ES	99457	IO	187213
12	A	77354	ER	153069	ST	94753	TE	181527
13	ON	74111	RO	150380	AL	94641	EN	178343
14	TI	72399	S	146605	IO	94227	AL	164986
15	AN	70483	IO	145203	AN	89316	F	164151
16	O	68994	TE	141872	RE	81527	OF	163853
17	2	67579	EN	140844	TR	80929	TH	163250
18	6	64101	P	140558	RO	78697	D	155882
19	T	60078	AL	135155	EN	78636	ES	153562
20	AT	58288	TH	127736	RA	78512	IC	147353
21	04	58029	T	122419	NG	73009	RE	142821
22	TH	57332	IC	122226	R	69385	NE	142567
23	2	55416	I	117330	ME	68806	HE	136844
24	ER	54707	D	117249	D	68681	RA	135704
25	0	54601	O	116693	CT	68191	L	131615
26	C	54498	HE	113895	OR	68177	OR	130380
27	S	52992	RE	113656	CO	66367	OL	122982
28	RA	52772	NE	112503	L	65937	TR	122825
29	HE	51347	RA	105842	RI	64187	CT	121237
30	F	50142	OR	104542	NT	63315	ST	121183
31	OF	49705	OL	98516	ND	60940	RI	119992
32	D	49360	RI	98208	LE	59537	ET	116967
33	IO	48383	ST	96639	EL	59537	DE	115967
34	15	47861	CO	96196	G	57761	CO	113005
35	01	47444	DE	94717	F	56872	ME	112330
36	P	46895	L	93237	OF	56725	NT	110573
37	EN	46613	CT	93118	SI	54755	LE	110485
38	AI	46223	ET	92509	HE	53824	IT	109639
39	ES	45958	TR	92277	T	53081	R	107113
40	8	45923	M	89217	EC	51840	ND	102621
41	ST	43445	NT	88791	TH	51688	IO	102111
42	02	43370	IT	87961	Y	51376	AR	96087

Figure 6-19

FREQUENCY-ORDERED BIGRAM LISTS

Bigram	CA Vol 75 Position	CA Vol 73 Position	EI Vol 71 Position	BA Vol 52 Position
Eb̸	1	5	7	4
Nb̸	2	3	5	7
IN	3	2	2	9
ON	4	4	3	13
Sb̸	5	1	1	5
TI	6	6	4	14
AT	7	9	8	20
AN	8	10	15	15
ER	9	12	6	24
RO	10	13	18	47
IO	11	15	14	33
TE	12	16	9	48
EN	13	17	19	37
AL	14	19	13	38
Fb̸	15	65	35	30

Table 6-6

DATA BASE BIGRAM COMPARISON

- for CA (based on Vol. 73)
  b̸EIOANTSRLCDMHPUYFGBV.X,K-WZ);QJ123('450978+/$=*? :
- for CA (based on Vol. 75)
  EIONb̸TARSLCDMUHPYFGBVXZWKQ21,34(0.56J79)8 '+:=?;*$/-

Just as the Least Common Bigram affects search time, (Section 5.6) so does the individual character frequency, though not to so great a degree. The SEARCH program executes the built-in function INDEX over a million times in an average run, and the time required for this execution is dependent upon the relative positions of characters in the look-up string. For maximum efficiency, these characters should be ordered in decreasing frequency order of single characters in the data base.

6.3 Data Base Terminology Variation

One of the problems associated with profile preparation is the use of identical terms in different data bases. Technically, a profile can be run against multiple data bases and will cause hits only in the data bases where the terms occur. Although it can be (and is) done, it is not the best method--the same term in multiple data bases can have different meanings or provide a different degree of specificity because of the nature of the file; for example, the term ACID as used in a chemical data base, an engineering data base, and sociological data base would function differently. In Chemical Abstracts, (Figure 6-20) it would be a non-specific term of high frequency (11,868 occurences in 1/2 year) that would have to be "AND"d to other terms. In Engineering Index the term ACID would be a reasonably specific low frequency term (See Figure 6-21)(253 occurences in 1/2 year) that might even stand alone as a search term. In a sociological data base the term ACID would probably refer to LSD.

Another set of examples of variation in terminology among data bases is given in Figure 6-21 where we see, for example, that proper names, compounds, formulas, isotopes, and Greek letters are represented differently in CA, BA, and EI.

	ACID	*ACID	ACID*	*ACID*	TOTAL
TERMS	1	31	55	43	130
TOTAL FREQUENCY IN 26 ISSUES	11868	72	3904	76	15920

Figure 6-20

ACID APPEARANCES IN CA

	ACID	*ACID	ACID*	*ACID*	TOTAL
TERMS	1	4	12	5	22
TOTAL FREQUENCY IN 10 ISSUES	423	5	129	5	562

Figure 6-21

ACID APPEARANCES IN EI

Data Base	Term Representation
	John Q. Public Jr. (proper name)
CA	PUBLIC JOHN QUINCY,JR
CA	PUBLIC JOHN Q,JR
CA	PUBLIC J Q,JR
BA	PUBLIC JQ
EI	PUBLIC,JR JQ
	Lipoprotein (compound type)
CA	LIPOPROTEIN
BA	LIPO PROTEIN
	New York (city)
CA	NEW YORK
BA	NEW-YORK
	Sulfuric Acid (H_2SO_4) (formula)
CA	H2SO4
EI	H//2 SO//4
	Carbon 12 (isotope)
CA	CARBON-12-LABELLED
CA	C-12-LABELLED
CA	CARBON 12
CA	C 12
EI	**1**2C
EI	CARBON 12
	Alpha (Greek letter)
CA	.ALPHA.
BA	ALPHA

Figure 6-22

VARIATION IN TERM REPRESENTATION

7. USER AIDS

In order to assist the user in writing and monitoring his profile, including selection and truncation of terms, CSC has prepared a number of user aids in the form of documents, computer generated lists, and output card information.

The CSC Search Manual explains the basic techniques of profile writing. A Supplemental Guide has been written for each data base. The guide demonstrates profile writing tailored to the specific data base. A Truncation Guide illustrates where to truncate a term in order to retrieve the maximum relevant words with the minimum noise. For example, Figure 7-2 from the Truncation Guide demonstrates the retrieval ability of various forms of terms related to the concept "analysis."

Frequency Lists in Frequency Order and Frequency Lists in Alphabetic Order are prepared for each data base. (See Figures 6-10 and 6-7). These lists are used as rough indicators of the volume of output one might expect to receive for specific terms. They are prepared for one volume at a time for each data base and are updated periodically.

Key-Letter-in-Context (KLIC) indexes are prepared for each data base. The KLIC indexes indicate where letter combinations occur. They are used in conjunction with our Bigram Frequency lists which provide a frequency count for every two-letter combination(bigram) in the data base.

As further aids to users in monitoring their output, Index Terms and Hit Terms are printed on each output card (see Figure 3-2) to provide the user with information for revising his profiles; Search Term Frequency/Issue listings are generated for each profile to show the user the frequency of occurrence in the issue searched for each term in his profile.

7.1 Search Manual and Supplemental Guide

In preparation for user education, workshops, and training seminars, IITRI developed a Search Manual. The manual was designed to assist CSC users in developing individualized search profiles for use with the CSC system. In preparing a profile the user prepares the detailed specifications he requires for retrieving citations from a data base. The manual explains the problems and techniques associated with development of search profiles. Problem areas include: the inflexibility of machinable data bases; the variety of word forms (grammatic, semantic, syntactic, and generic); the variety of conventions employed for abbreviations, symbols, and acronyms; the varied practices, degrees of specification, and presence or absence of controls employed in indexing and classification; and the variety of nomenclature used within and among data bases.

The special techniques of profile preparation are: determination of search terms--including synonyms, higher and lower generic terms, and related terms; determination of searchable entries other than subject terms, such as authors; the use of left and right truncation for retrieval based on term fragments and distinctive letter combinations; the use of links for grouping of related terms within a logic expression; development of free-form logic expressions employing the Boolean operators AND, OR, and NOT; and the assignment of weights to profile terms in accordance with relative importance of terms to the user.

A Supplemental Guide has been prepared for each data base searched. The guide provides information about the use of data elements that are specific to the particular data base and demonstrates profile writing techniques for that data base.

7.2 Key-Letter-in-Context (KLIC) Indexes

Key-Letter-in-Context (KLIC) indexes* are prepared for each data base to assist users in selecting term fragments. The KLIC index is prepared from title and keyword terms appearing in the data base. A KLIC index is similar to a Key-Word-in-Context (KWIC) index but is confined to a single term and alphabetizes the term separately under each of its constituent characters indicating preceding and following characters as they are wrapped around the distinguishing character. The KLIC index is a lexicographic ordering of terms in a data base by each character (alpha, numeric, or special) in the term or character string. It is a permuted term arrangement sorted by character. The format of a KLIC index is shown in Figure 6-15.

---

*We are indebted to Dr. Anthony Kent of the University of Nottingham for the concept of the KLIC index and the insight into its utility.

A user can see what the potential retrieval may be from using a term fragment in any of the truncation modes. The KLIC index is especially helpful in selecting fragments with left truncation or both left and right truncation.

A program, KLICPT, prints the KLIC index of all the words in a four column format with the frequency of the term following the term as shown in Figure 6-15. Delimiters for terms in the KLIC index are asterisks, slashes, and blanks. Each entry in the index appears in a 21-character line, with the eleventh character as the sort character. A double slash (//) is used as a word delimiter, and the words are wrapped around the central sort character. The KLIC index is used for linguistic research and as a user aid. By consulting the KLIC index one can determine the retrieval capability of a particular letter combination or term fragment. The KLIC index is used to identify letter combinations that are highly specific and would therefore be discriminating search terms, e.g., the character string *YBD* does not occur anywhere in the CA or BA data bases except in the term MOLYBDENUM (Note: in a literary data base it would occur in the mythological characters SCYLLA and CHARYBDIS). Thus, *YBD* could be used as a search term for molybdenum. On the other hand, letter combinations that occur frequently in many irrelevant terms should be avoided, e.g., the letters RNA for ribonucleic acid could be used as a search term assuming one did not specify simultaneous left and right truncation *RNA*. The simultaneous truncation mode would retrieve more than 200 irrelevant words. Some of these are:

ALTERNATE
BARNACLE
CARNATION
DIURNAL
FINGERNAIL
MATERNAL

From Figure 6-15 it can be seen that a user who employs the search term *ACID might expect different terms to be retrieved. *ACID* might retrieve terms that include several cases of singular and plural word forms. The KLIC Index can only be used as a general guide, as the terms appearing in any given issue of a data base will not necessarily correspond with the list appearing in the index.

Also the term fraction occurrences for a given term in different data bases will differ, hence searching on the same truncated word in different data bases will retrieve different terms.

7.3 Term Frequency Lists

Frequency Lists in Frequency Order and Frequency Lists in Alphabetic Order have been prepared for each data base. They are used to assist in selection of search terms. Figure 6-10 shows a portion of a frequency-ordered term frequency list and Figure 6-7 shows a portion of the alphabetically-ordered term frequency list. A high frequency term will produce a high volume of hits unless it is combined with another search term or assigned a low weight. For this reason we have instituted an automatic check to notify us if profiles contain any of the 50 highest frequency terms in a given data base. If any of these terms are used they must be AND'd to other terms, assigned a low weight, or otherwise restricted within a profile. A low frequency word might be used independently. Frequency lists are used as rough indicators of the volume of output one might expect to receive for specific terms. Our frequency lists have been prepared for one volume of each data base.

7.4 Truncation Guide and Standard Truncation

Truncation is used to facilitate retrieval of terms containing fragments that are common to two or more different forms of a term. The use of a single fragment will retrieve all terms containing that fragment in accordance with one of the truncation modes as described in Section 4.2.2

Individual users were initially allowed to use search terms and truncations in a free and uncontrolled manner. A study of the resulting aggregated term list revealed numerous sets of related term fragments. Examples of related fragment sets are shown in Figure 7-1.

For commonly used terms a preferred truncation can be selected that will meet three conditions:

(1) The truncated term is a fragment common to a set of desired words associated with that fragment.

(2) The fragment is unique and its use will not retrieve other terms outside of the associated set.

(3) The fragment is the shortest representation that preserves uniqueness.

PREPAR*	PURIF*	SYNTH*
PREPARAT*	PURIFIC*	SYNTHE*
PREPARATION	PURIFICAT*	SYNTHES*
PREPARATION*	PURIFICATION	SYNTHESI*
	PURIFY	SYNTHESIS
PREPAR*	PURIF*	SYNTHE*

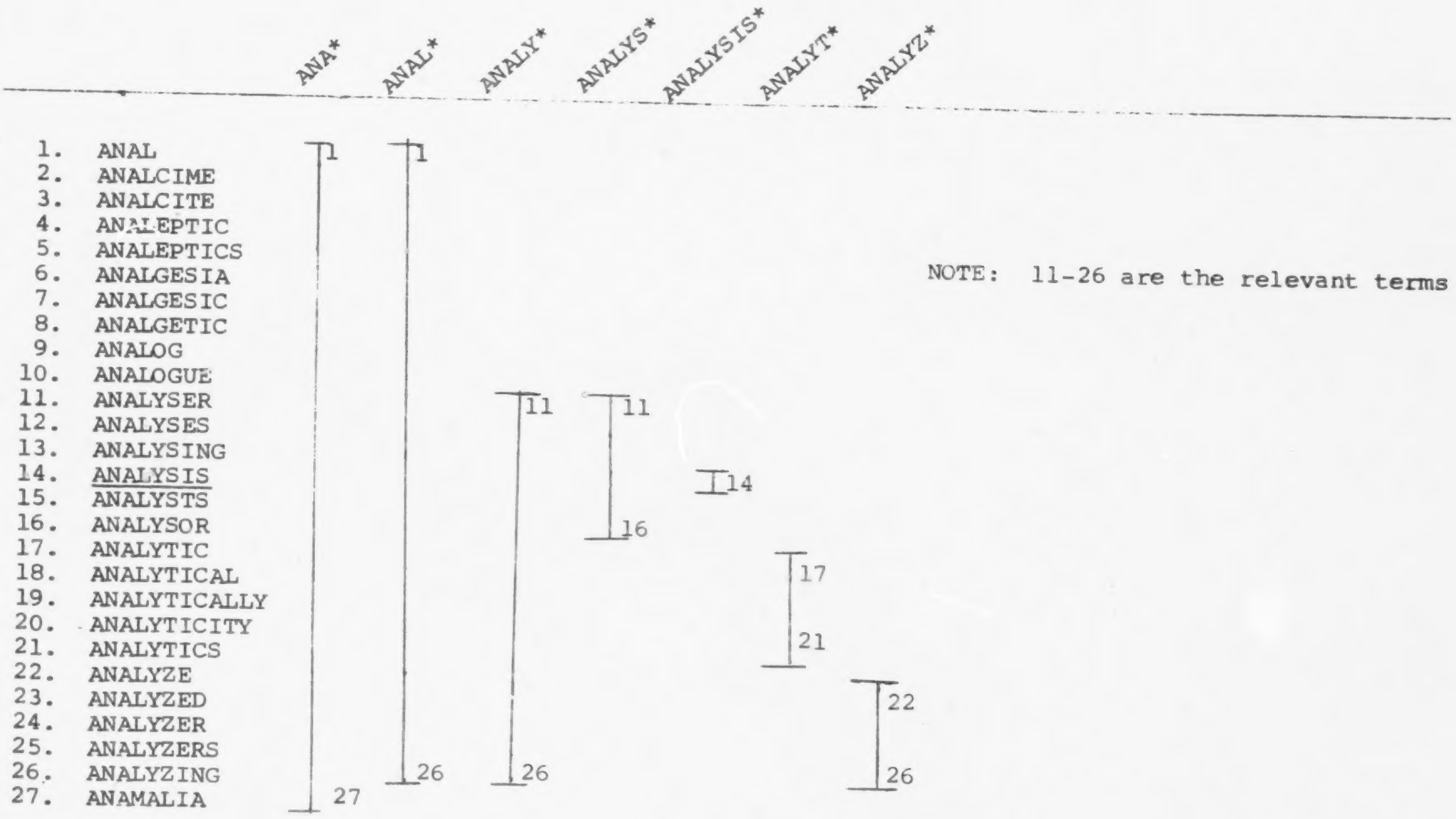
*denotes truncation; fragment below line in each set is the preferred truncation.

Figure 7-1
SETS OF RELATED FRAGMENTS WITH UNCONTROLLED TRUNCATION

Inasmuch as CSC search time is directly proportional to the number of profile terms in a run, we can reduce search time by establishing standard truncations for concepts used by several users. In one of our early runs we found that in a number of instances the variety of truncation forms used was significant. Accordingly, a Truncation Guide has been prepared that lists more than 600 fragments of which 151 have been recommended for use. The truncations listed in the Guide are all right truncations (mode 2) and were selected from a list of common terms that have been employed by users of CSC. Each term has been placed in alphabetical order within a set of words associated with varying length term fragments. The words for the alphabetical listings were obtained from Chemical Abstracts Service Search Guide; The Condensed Chemical Dictionary, 5th edition; Webster's Seventh New Collegiate Dictionary; and Chemical Abstracts Index. A page from the Truncation Guide is shown in Figure 7-2 .

At the top of each listing appears a set of candidate truncations or term fragments. The brackets in each column identify the terms in the alphabetical list that would be retrieved by the use of the designated fragment. A term fragment is considered optimal if it satisfies all three of the conditions stated above. Other fragments may provide either over-truncation or under-truncation. In over-truncation, the fragment is too short and an overlapping of more than one set of associated words occurs leading to the retrieval of non-relevant terms. In under-truncation, the fragment is too long and a loss of relevant terms may occur due to the excessive restriction on the set of terms that can be retrieved. In some cases, several fragments of varying lengths will retrieve the same set of terms. The shortest fragment is then selected as optimal.

While the CSC Truncation Guide is helpful, one can achieve many of the same objectives by using a handbook, dictionary, or other list of terms. They can be used:


ANA*
ANAL*
ANALY*
ANALYS*
ANALYSIS*
ANALYT*
ANALYZ*
1. ANAL
2. ANALCIME
3. ANALCITE
4. ANALEPTIC
5. ANALEPTICS
6. ANALGESIA
7. ANALGESIC
8. ANALGETIC
9. ANALOG
10. ANALOGUE
11. ANALYSER
12. ANALYSES
13. ANALYSING
14. ANALYSIS
15. ANALYSTS
16. ANALYSOR
17. ANALYTIC
18. ANALYTICAL
19. ANALYTICALLY
20. ANALYTICITY
21. ANALYTICS
22. ANALYZE
23. ANALYZED
24. ANALYZER
25. ANALYZERS
26. ANALYZING
27. ANAMALIA
1
27
1
26
11
26
11
16
14
17
21
22
26

NOTE: 11-26 are the relevant terms

Figure 7- 2

TRUNCATION GUIDE ENTRIES FOR THE CONCEPT "ANALYSIS"

(1) To obtain an estimate of the number of discrete terms and type of terms that may be retrieved by using right truncation with a given term fragment.

(2) To balance selection between a longer and shorter term fragment.

(3) To indicate optimal term fragments that are the shortest, unique fragment capable of retrieving a set of associated words.

(4) To designate fragments for use with terms where a seemingly optimal truncation may be ambiguous or lead to false retrieval.

Standard truncations as indicated in the Truncation Guide are used not only because they improve retrieval effectiveness--they also provide a cost savings to CSC as their use increases the aggretation ratio for profile input terms. A check of several groups of profiles indicated that 10% of the terms could employ standard truncations. However, it is not always possible to use the optimal truncated form of a word as a standard form. In some cases, the data base may contain abbreviations that are different from the optimal truncation forms of a full word and the abbreviation must be used to ensure retrieval. For example, in the case of the concept ANALYSIS, CA uses the appreviation ANAL for the group of words which we have determined to be best found with the optimal truncation form ANALY*. Both terms should be used, ANALY* to retrieve from text, and ANAL (no truncation) to retrieve from the CA keyword list. In other cases, truncated words may retrieve too many false drops. For example *AMIN* will retrieve various amines, but it will also pick up words such as CONTAMINATION. In another case one user may be interested in crystals and all forms of the term.

He would use the truncated term CRYSTAL*. Another user may wish only the process crystallization and not everything on crystals or crystallography. Such a user would use CRYSTALLIZ* rather than CRYSTAL*. The profile preparer can not blindly select truncation forms from the Truncation Guide or other aids. Each selection must be made in full understanding of the profile and the data base.

8. USER EVALUATION AND FEEDBACK

For purposes of assessing the degree of user satisfaction and usefulness of the SDI, system users have been requested to evaluate the relevance of retrieved citations. A relevant citation is defined as one that is judged by the user to satisfy the intent of his profile. Although the output citation necessarily satisfied the search terms, logic, and associated parameters of the profile inasmuch as they were the keys that retrieved the citation, the intent of the profile may not be necessarily satisfied. It is the user's judgment that is required to discern the discrepancy between intent and output and modify his profile accordingly.

Evaluating or attempting to measure the performance, effectiveness, and utility of an information retrieval system is difficult for a variety of reasons not the least of which is that the user's interests may change over a period of time. An article that is of interest today may not be of great interest to him a month hence and vice versa. Because user interests and profiles change, we have requested that users evaluate citations at, or as close as possible to, the time of receipt from the Center.

An evaluation form (shown in Figure 8-1) is sent to each user with every issue of output. The evaluation form indicates the number of citations sent for the particular SDI run and asks the user how many of these were of interest, and of no interest.

Using the values returned on the evaluation reports, the CSC calculates the percent relevance (precision) of output for each user. Data are accumulated by user, by company, and by issue. When these reports indicate that the user is receiving too much extraneous material or very little pertinent output, the CSC personnel consult the individual user with suggestions and assistance if modifications are needed.

If a user's precision rating runs high or low over several runs this usually indicates a problem. If he gets 90%-100%

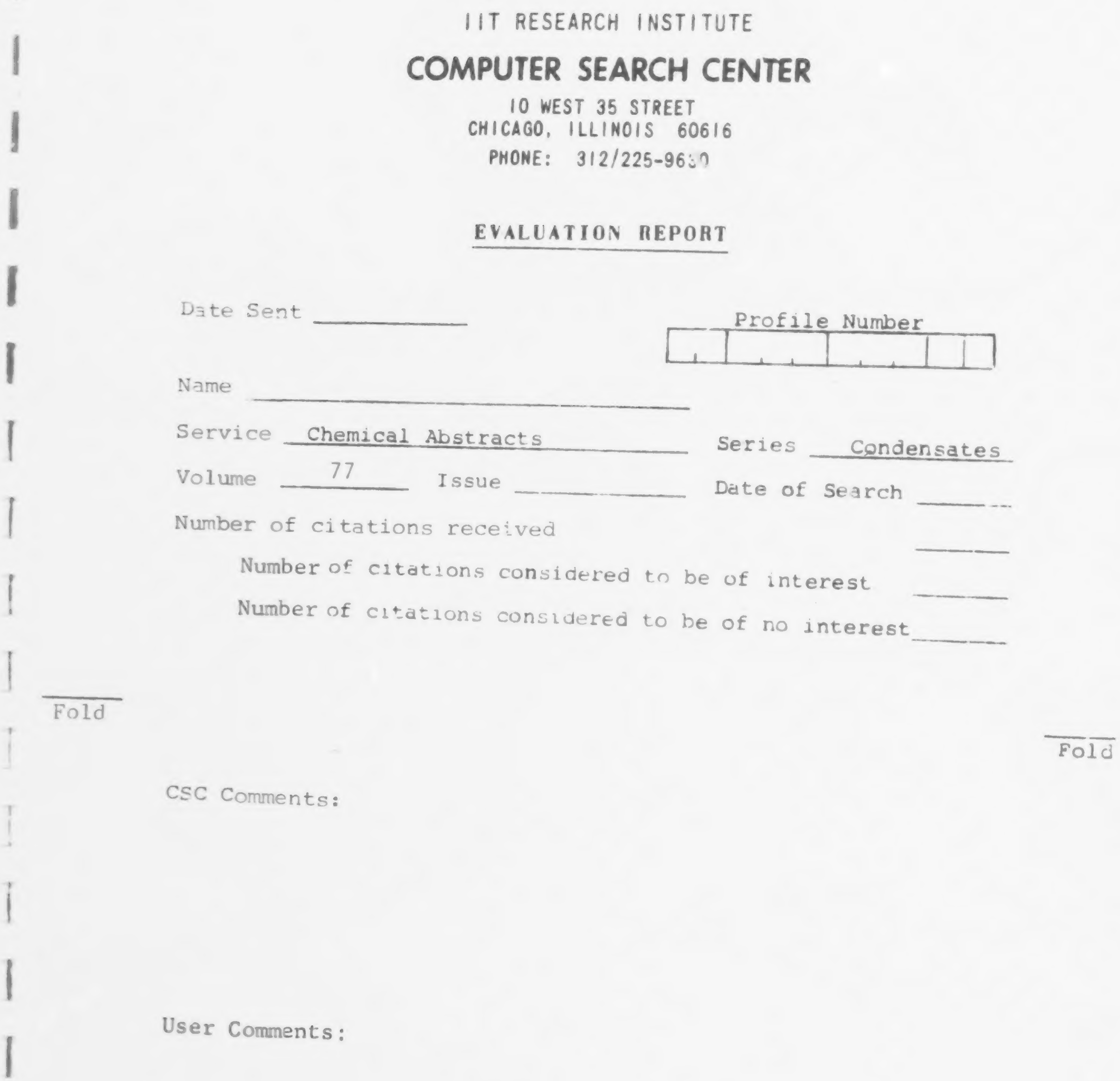
precision he is probably missing relevant citations by using terms that are highly specific or logic that is overly tight. If he gets precision ratings below 25%,he is getting too many non-relevant citations and this is probably due to the use of high frequency or common terms in an unrestrictive manner.

A high percentage of the forms are returned (see Table 8-1) which indicates that the users are looking at their output and checking it--at least to the extent of putting the cards in the two fill-in groups: relevant and non-relevant. In general, 50 percent of the forms are returned to IITRI within two weeks of our mailing. The balance of the forms are returned anywhere from three weeks to ten months after the mailing.

Precision was calculated as the number of citations considered to be of interest by the user divided by the number of citations sent to the user as indicated on the returned evaluation reports. The statistics also do not consider the fact that when no citations were located,zero output might well represent real information and in effect be 100 percent satisfactory to the user.

Precision statistics are presented in Table 8-2. The statistics were obtained from 131 searches run on CA Condensates from Volume 71, issue 9 through Volume 76,issue 12.

Table 8-2 lists average precision ratings of retrieved citations by search run. These numbers are affected by the content of CA Condensates. Content varies from week to week since not all journals are abstracted in every weekly issue. Some profiles would therefore be low in citations of interest retrieved in a week when the journals in the area of their interest are not abstracted. The figures for the 131 weeks listed in Table 8-2 vary from a low of 19.0 percent to a top figure of 46.5 percent with an average weekly relevance of 30.0 percent. The weekly average was calculated by averaging the percent relevance of the individual users. The



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

EVALUATION REPORT

Date Sent

Profile Number

Name

Service Chemical Abstracts

Series Condensates

Volume 77

Issue

Date of Search

Number of citations received

Number of citations considered to be of interest

Number of citations considered to be of no interest

Fold

Fold

CSC Comments:

User Comments:

Figure 8-1
USER EVALUATION REPORT FORM

CHEMICAL ABSTRACTS CONDENSATES

ISSUE	VOLUME 71	VOLUME 72	VOLUME 73	VOLUME 74	VOLUME 75	VOLUME 76
1	(Data for issue nos. 1-8 does not exist)	89.9	85.1	84.9	67.6	53.6
2		94.8	88.3	75.8	68.0	51.1
3		88.5	82.1	84.6	67.9	55.4
4		90.7	86.7	78.3	67.3	56.0
5		91.8	84.1	85.2	65.3	52.9
6		91.4	87.7	80.0	62.0	44.1
7		90.9	79.4	75.0	61.2	53.8
8		94.7	88.0	69.4	64.8	45.3
9	89.0	90.0	78.5	78.5	61.1	48.9
10	96.9	90.2	76.5	70.0	66.0	34.9
11	90.0	86.8	82.3	74.2	61.3	49.5
12	88.4	88.5	80.7	71.4	65.5	45.5
13	93.3	87.5	76.5	75.0	59.6	
14	95.2	85.5	80.4	69.0	62.8	
15	84.6	81.3	77.9	72.7	56.9	
16	92.8	90.3	76.7	67.4	58.0	
17	91.0	83.7	84.4	65.9	56.1	
18	94.8	82.8	81.0	69.0	56.8	
19	91.4	73.1	86.6	63.5	54.1	
20	94.7	70.7	77.1	66.7	56.1	
21	92.3	79.0	81.7	67.4	63.5	
22	91.8	68.5	80.9	65.3	55.9	
23	93.6	73.9	82.1	71.3	57.6	
24	(Data for issue nos. 24-26 does not exist)	73.4	78.9	57.7	55.7	
25		73.6	78.4	61.1	54.2	
26		74.6	72.2	59.1	53.5	
AVG	92.0	84.1	81.3	71.5	60.7	49.2

Table 8-1

RETURN OF EVALUATION FORMS VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES

ISSUE	VOLUME 71	VOLUME 72	VOLUME 73	VOLUME 74	VOLUME 75	VOLUME 76
1	(Data for issue nos. 1-8 does not exist)	32.9	31.5	26.9	29.5	28.9
2		34.2	32.3	30.0	34.0	28.6
3		32.0	24.6	24.3	25.5	29.9
4		28.2	28.9	29.0	30.9	27.6
5		30.3	27.6	24.8	28.0	27.2
6		32.1	31.0	26.8	33.3	34.5
7		33.3	24.0	23.0	29.2	26.7
8		29.1	26.1	19.0	31.3	24.9
9	46.5	31.1	23.6	24.2	30.0	21.5
10	37.1	26.5	31.8	27.6	35.0	34.3
11	37.3	40.3	22.8	25.2	27.1	28.1
12	43.4	24.1	29.3	28.6	30.0	32.3
13	42.6	35.3	22.1	27.5	26.2	
14	43.3	28.5	29.0	29.0	33.4	
15	41.3	33.6	26.1	27.1	26.6	
16	35.6	25.9	25.9	27.2	33.8	
17	42.1	34.4	28.5	31.0	24.6	
18	33.3	30.1	25.5	24.4	35.1	
19	37.4	34.9	24.1	28.3	29.6	
20	25.9	32.5	30.9	33.0	32.1	
21	33.8	29.5	25.1	27.8	28.3	
22	32.5	26.5	34.2	28.1	34.9	
23	32.0	29.2	24.6	27.9	27.1	
24	(Data for issue nos. 24-26 does not exist)	27.0	28.5	33.8	33.9	
25		25.3	25.5	32.3	28.6	
26		27.3	31.6	27.5	33.8	
AVG	37.6	30.5	27.5	27.5	30.5	28.7

Table 8-2
PRECISION VS. ISSUE

figures vary not only because of the availability of material on the particular question but also because of the attitude of the user toward modifications of his profile. In many cases, modifications have been made by CSC personnel and the users cooperatively. In other cases, users have taken the initiative in modifying their own profiles. But there are cases where the user has not wished to modify his profile and in these cases, it is possible that citations that could be pertinent are being missed or too much irrelevant material is being produced. This situation does cause some low relevance ratings for individual users.

The distribution of the average profile precision for profiles that were searched in all of the 131 runs of CA Condensates Volume 71, issue 9 through Volume 76, issue 12 is shown in Table 8-3. Average precision for 12 issues CA Volume 76 and 15 issues CA Volume 71 ranged from 0 to 100 percent. More than 50 percent had greater than 50 percent relevance.

Percent Relevance	CA Volume 76 (12 issues) Percent Profiles	CA Volume 71 (15 issues) Percent Profiles
0	16.4	28.8
1-10	14.9	4.9
11-20	13.4	7.2
21-30	8.8	9.2
31-40	8.9	8.1
41-50	9.3	15.6
51-60	5.0	4.7
61-70	5.9	4.9
71-80	3.7	4.3
81-90	3.6	1.8
91-100	10.1	14.9

Table 8-3
DISTRIBUTION OF AVERAGE PROFILE PRECISION

In addition to the evaluation forms for monitoring precision ratings CSC requests users to send back the trailer card (see Figure 3-3) from their output after circling the citation numbers for the relevant citations. In this way the CSC profile coordinator can see exactly which citations are considered to be of interest. This helps her to understand the user's interest so that she can suggest more meaningful profile changes.

In addition to precision data obtained on a weekly basis throughout the program, CSC carried out a study to obtain more detailed information and evaluations from its users. In mid-June 1970 a questionnaire, User Evaluation of Current Awareness Service for Chemical Abstracts Condensates, was sent to all current users.

Table 8-4 is a summary of responses of 51 users of CSC SDI system searching CA Condensates for 71 profiles.

QUESTION		IIT/IITRI	OTHER ACADEMIC	IND.	TOTAL
1 CA available	yes	9	11	42	62
	no	-	-	-	-
2 Prior manual search	yes	4	6	23	33
	no	5	5	18	28
3 Monitor searches	yes	5	4	23	32
	no	4	6	18	28
4 Dispense with manual searches	yes	6	3	22	31
	no	1	7	17	25
5a Card format satisfactory	yes	8	8	37	53
	no	1	-	3	4
5b Index terms	Useful	9	7	36	52
	Not	-	2	6	8
5c Terms causing hits	Useful	8	7	33	48
	Not	1	1	9	11
6 Maintain card file	yes	9	6	32	47
	no	1	4	9	14
7 Card file useful	yes	9	4	22	35
	no	-	1	10	11
8 Look up citations	yes	8	9	40	57
	no	1	1	1	3
9 Hard-copy retrieval	Personal	8	6	26	42
	Librarian	-	3	22	25
12 Modifications could improve profile	yes	5	7	28	40
	no	4	2	13	19
13 Distribution of cards prompt	yes	9	10	41	60
	no	-	-	-	-
14 Profile liaison	Sat.	9	9	34	52
	Unsat.	-	-	4	4
15 Subscription desirable	yes	6	3	32	41
	no	-	5	5	10

Table 8-4

SUMMARY
USER EVALUATION OF CA CONDENSATES
CURRENT AWARENESS SERVICE

9. EDUCATION--USER LIAISON

One problem facing information centers is that of education. The machine-readable sources of information are not familiar to the average working scientist. In order to familiarize the potential users of information centers with the new sources and services, we at IITRI have undertaken a number of educational activities. Education must pave the way for marketing. The means we have undertaken include: development and/or conduct of workshops, seminars, university courses, short courses, workbooks, technical presentations, publications, and mass mailings. Once a user has entered one or more profiles in our system, it is necessary to maintain liaison with him for modification of his profile as changes occur in the data bases and/or his interests. Both aspects of center-user interaction are described in this section, that of basic education in the utility of SDI services from machine-readable data bases and that of continuing liaison while servicing a profile.

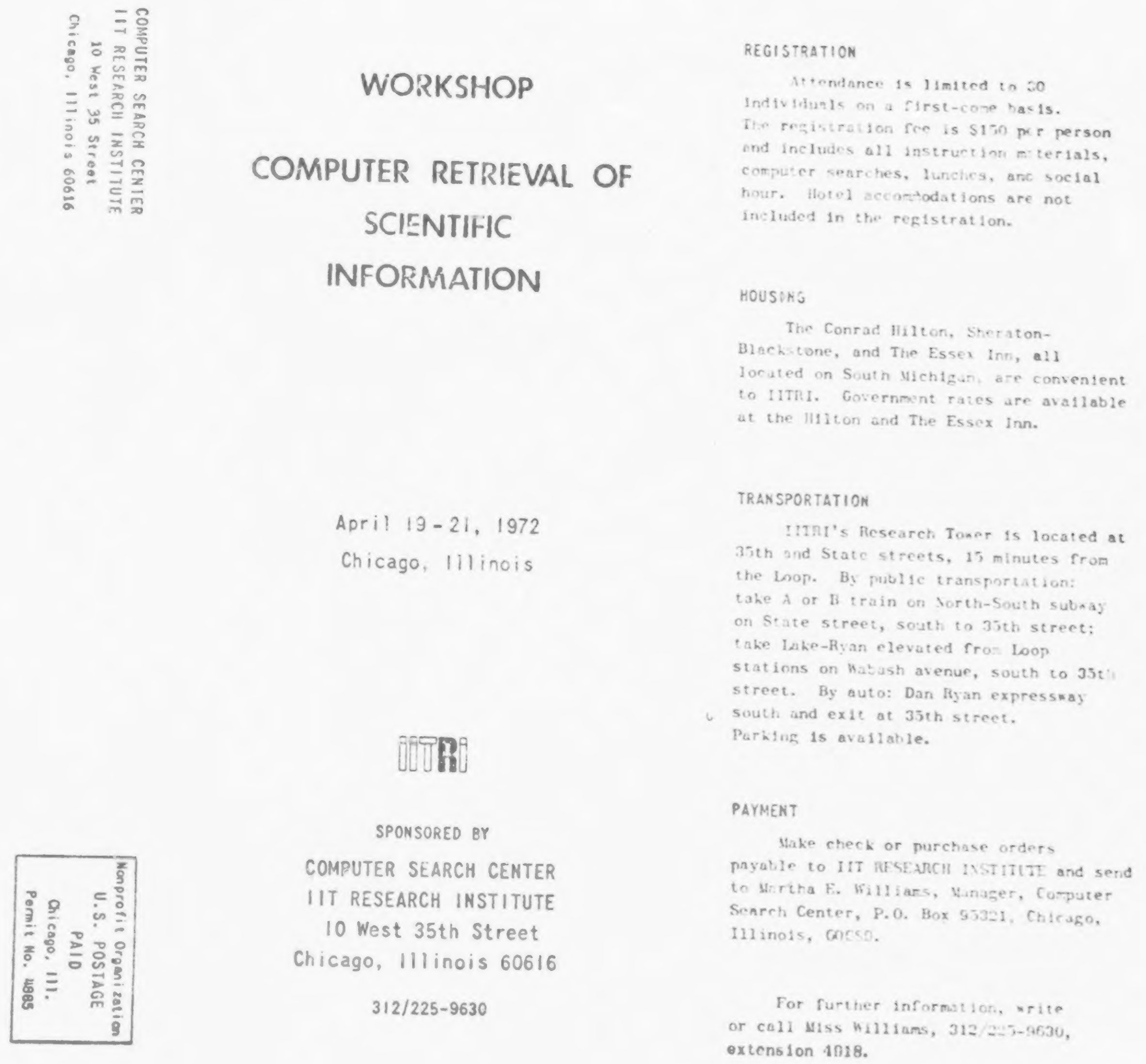
The educational aspects of our workshops, seminars, etc., are devoted to providing basic information on machine-readable data bases and their use, in terms of data base contents and limitations, machine search capabilities and limitations and the advantages of mechanized SDI service. There are many advantages to using SDI services of information centers. Our system was designed to provide many advantages and through the past three and a half years of operating experience we have both become aware of more advantages and gained considerable data to substantiate our original assumptions. The most obvious reasons for using SDI services include: (1) coverage, (2) thoroughness of search, (3) consistency of search, (4) interdisciplinariness, (5) high recall, (6) cost-effectiveness, (7) speed and regularity, (8) timeliness, (9) multiplicity of data bases, (10) automatic preparation of files in standardized format, and (11) cost of data base preparation and operation of an SDI system vs. subscriptions. Further details on these eleven items are presented in a paper entitled "Handling of Varied Data Bases in an Information Center Environment" published in the Proceedings of the Conference on Computers in Chemical Education and Research, Northern Illinois University, DeKalb, Illinois, July 19-23, 1971.[5]

9.1 Workshops on Computer Retrieval of Scientific Information

We have conducted four workshops for industrial and academic participants in the use of computer techniques for retrieval of scientific information. They were held on January 19-21, 1971, May 5-7, 1971, Dec. 1-3, 1971 and April 19-21, 1972. Another is planned for November or December of 1972. Each of these Workshops consists of an intensive 2½-day program of lectures and "hands-on" use of the CSC's SDI service. Figures 9-1 and 9-2 show the front and back of the Workshop announcement brochure that is mailed to prospective participants. CSC staff members give lectures on: CSC philosophy and operations; techniques for preparing search profiles including use of data elements, truncation, links, logic, and weights; the characteristics of data bases; use of aids such as frequency lists, KLIC (Key-Letter-in-Context) indexes, and truncation guides; theory of retrieval evaluation including recall, precision and feedback; and on modification of search questions.

Attendees write profiles to reflect their areas of interest. Profiles are run against representative issues of CA Condensates, BA Previews, and/or EI COMPENDEX. Following the first run attendees conduct manual searches of the appropriate hard copies of CA, BA, and/or EI to compare the results of the machine search against manual searches. Profiles are then evaluated and modified and submitted for a second machine search against the same data bases. Output from the second run is also evaluated by attendees.

Figure 9-3 presents data on recall and precision taken from the CA searches made by participants of the third Workshop. Since both manual and machine searches are made, it is possible to calculate both recall and precision. The increase in both of these indicators after profile revision has been observed in all Workshops.



COMPUTER SEARCH CENTER
IIT RESEARCH INSTITUTE
10 West 35 Street
Chicago, Illinois 60616

WORKSHOP

COMPUTER RETRIEVAL OF
SCIENTIFIC
INFORMATION

April 19-21, 1972
Chicago, Illinois

IITRI

SPONSORED BY

COMPUTER SEARCH CENTER
IIT RESEARCH INSTITUTE
10 West 35th Street
Chicago, Illinois 60616

312/225-9630

Nonprofit Organization
U.S. POSTAGE
PAID
Chicago, Ill.
Permit No. 4885

REGISTRATION

Attendance is limited to 30 individuals on a first-come basis. The registration fee is $150 per person and includes all instruction materials, computer searches, lunches, and social hour. Hotel accommodations are not included in the registration.

HOUSING

The Conrad Hilton, Sheraton-Blackstone, and The Essex Inn, all located on South Michigan, are convenient to IITRI. Government rates are available at the Hilton and The Essex Inn.

TRANSPORTATION

IITRI's Research Tower is located at 35th and State streets, 15 minutes from the Loop. By public transportation: take A or B train on North-South subway on State street, south to 35th street; take Lake-Ryan elevated from Loop stations on Wabash avenue, south to 35th street. By auto: Dan Ryan expressway south and exit at 35th street. Parking is available.

PAYMENT

Make check or purchase orders payable to IIT RESEARCH INSTITUTE and send to Martha E. Williams, Manager, Computer Search Center, P.O. Box 95321, Chicago, Illinois, 60690.

For further information, write or call Miss Williams, 312/225-9630, extension 4018.

Figure 9-1

WORKSHOP ANNOUNCEMENT BROCHURE--FRONT SIDE

WORKSHOP

COMPUTER RETRIEVAL OF SCIENTIFIC INFORMATION

The Workshop is a 2½ day program to acquaint attendees with the fundamentals of searching scientific literature with a computer. The philosophy and operations of the Computer Search Center will be described and the characteristics of major data bases will be discussed.

Each registrant should be prepared with search questions so that he or she can write realistic profiles. Sessions on profile development will culminate in the preparation of coded profiles that will be submitted to the Center for searching against current data tapes. Outputs will be analyzed and evaluated, profiles will be modified and a second search will be made to provide comparative data for evaluation.

The Workshop is designed for information specialists, technical librarians, and research personnel from government, industry, and academic institutions. Individuals who use technical literature or who are responsible for a firm's access to the technical literature will have a firsthand opportunity to participate in a program that can assist them in their information retrieval and dissemination tasks.

PROGRAM

FIRST DAY - WEDNESDAY

8:30 - 9:00 AM
Registration

9:00 - 12:00 AM
Introduction to Computer Searching
Indexing, Coding, and Computer Handling of Information
Computer Search Center - Overview
Characteristics of Machine Readable Data Bases

12:00 - 1:00 PM
Lunch

1:00 - 4:30 PM
Profile Writing Techniques
User Aids
Profile writing session -- participants will develop a search profile using the SEARCH Manual and other guides and will code the profile for computer searching of the selected data base(s)

SECOND DAY - THURSDAY

9:00 - 12:00 AM
Theory of Retrieval Evaluation
Evaluation of retrieved output: check search output against profile and printed issues of abstract journals
Modification of profiles

12:00 - 1:00 PM
Lunch

1:00 - 4:30 PM
Calculation of Precision and Recall
Evaluation Summary
Other services: personal libraries; multiple outputs; hardcopy and microfilm; group profiles, specialized data bases; survey of data bases

5:00 - 6:00 PM
Social Hour

THIRD DAY - FRIDAY

9:00 - 12:00 AM
Evaluation of modified profiles
Future activities in the field of information science

APPLICATION FOR ATTENDANCE

I plan to attend the Workshop

NAME ______________________________
FIRM ______________________________
STREET ______________________________
CITY ____________ STATE ____________ ZIP ____________
TITLE ______________________________
TELEPHONE AREA ______ /NUMBER ______________
I AM A CITIZEN OF ______________________
MY PRIMARY FIELD OF INTEREST IS ______________________

Return to: MARTHA E. WILLIAMS, COMPUTER SEARCH CENTER,
IIT RESEARCH INSTITUTE, P.O. BOX 95321,
CHICAGO, ILLINOIS 60680

☐ CHECK ENCLOSED
☐ PURCHASE ORDER ENCLOSED

Figure 9-2

WORKSHOP ANNOUNCEMENT BROCHURE--REVERSE SIDE

December 1, 1971

	MANUAL	MACHINE				
Profile Number	Cits. Ret'd	Cits. Ret'd	Relevant	Total Rel.	Recall	Precision
001-1	5	9	3	5	60	75
001-2	0	0	0	0	N/A	N/A
001-3	2	0	0	2	0	N/A
002-1	6	3	2	7	29	67
002-2	8	14	11	11	100	79
003-1	8	6	4	8	50	67
004-1	4	13	4	4	100	30
005-1	0	0	0	0	0	0
005-2	4	6	2	6	67	33
006-1	1	1	1	1	100	100
006-2	0	2	0	0	N/A	0
007-1	10	3	1	10	10	33
008-1						
009-1	2	1	1	2	50	100
009-2	3	3	3	3	100	100
010-1	1	13	1	1	100	8
010-2	1	7	1	1	100	13
011-1	14	21	11	16	69	52
012-1	74	25	7	80	9	28
			Average		59%	49%
		December 2, 1971				
001-1	5	8	7	7	100	88
001-2	2	2	2	2	100	100
002-1	8	8	5	11	73	62
002-2	14	22	18	18	100	82
003-1						
004-1	4	5	4	4	100	80
005-2	4	6	3	7	43	50
005-3	14	16	10	19	53	63
006-3	14	14	9	17	53	64
007-1						
007-2	3	12	11	13	84	91
008-2						
009-3	1	3	1	1	100	33
010-1	1	6	1	1	100	17
010-2	1	4	1	1	100	25
010-3	7	9	9	9	56	100
011-1	16	12	9	16	56	75
011-2	2	1	1	2	50	100
011-3	16	15	6	6	100	40
012-1						
013-1						
013-2						
			Average		79%	67%

Figure 9-3

WORKSHOP PRECISION AND RECALL STATISTICS

We have found the 2½-day period with opportunity for two manual and two machine searches to be a good format for presentation of this material. We have limited attendance to 20 to 30 people since we have found that individual instruction yields the best results. Although the workshops have proved to be a good source for continuing subscribers, they have frequently been attended by representatives of organizations that plan to implement their own system. Figure 9-4 presents a tabulation of the affiliations of attendees at the four Workshops. Those in the industrial area constitute more than 40% of the total participants.

9.2 Seminars

We have also conducted no-fee seminars as well as the more highly-structured Workshops. Seminars are two to four hours in length and are comprised primarily of the lecture portion of the Workshop material. Seminars are usually held for an individual company or university and will be conducted either at IITRI or on-site at the organization, depending upon their wishes. A large number of such seminars have been held and are listed in Section 10 of this report.

A general type of seminar on the CSC has also been presented as a case study within the framework of workshops conducted by the National Federation of Scientific Indexing and Abstracting Services (NFSAIS). We have presented this case study at NFSAIS workshops in Cleveland, Chicago, and New York, and the next is planned for Houston in mid-October of 1972.

9.3 University Courses

One of the more significant educational efforts has been carried out in cooperation with Illinois Institute of Technology, the university with which IITRI is affiliated. During the 1969, 1970, 1971, and 1972 spring semesters a new course was offered at IIT, "Modern Techniques in Chemical Information." The course was made available to second year graduate and upper division undergraduate students in the Chemistry Department. This course replaced the traditional

Workshop	Academic	Industrial	U.S. or Canada Gov't.	Tape Supplier	Public Library or Foundation or Research Institute	Not Specified	Total
Jan. 19-21 1971	6	7	2	2	2	0	19
May 5-7 1971	7	14	8	0	1	2	32
Dec. 1-3 1971	0	4	5	0	1	0	10
April 19-21 1972	4	9	8	0	1	0	22
Total	17	34	23	2	5	2	83
Percent	20.5	41.0	27.7	2.4	6.0	2.4	100.0

Figure 9-4

AFFILIATION OF WORKSHOP ATTENDEES

chemical literature course and the chemistry graduate students were given the option of taking the Modern Techniques course in lieu of a second foreign language. One hundred percent of the graduate students opted for the course. Members of the IIT staff who serve on graduate advisory committees willingly accepted this change as a significant improvement in the formal training for the Ph.D degree. One of the reasons for enthusiastic acceptance of the course is that it presents a solid basis for the understanding and use of chemical information systems in the context of a 2-credit hour one-semester course.

The course was made available through a sub-contract from the IITRI Computer Search Center program to the Chemistry Department at IIT and the course was taught by Dr. Paul E. Fanta of IIT and Miss Martha E. Williams of IITRI.

The course covered techniques of storage, search and retrieval of chemical information. Specifically, it stressed the fact that chemical information exists in many different forms, both printed and machine-readable, and if the chemist is to make good use of the multiplicity of available data bases and collections, he must expand his horizons and be prepared to use the computerized files as well as the traditional collections. Information resources and methods of retrieval were considered from the viewpoint of information systems and the general problem was considered to be the retrieval of specific data from a data store.

Inasmuch as none of the available chemical literature textbooks provide adequate coverage of the modern techniques and sources of chemical information, staff members from both IITRI and IIT (Mr. Eugene S. Schwartz and Miss Martha E. Williams of IITRI and Dr. Paul Fanta of IIT) developed a syllabus and workbook for the course, "Modern Techniques." The objectives and contents of the book are described in the following section.

In addition to acquainting the student with the traditional and modern methods of handling information, each of the students participated in an SDI program. Instruction in profile preparation was provided both through lectures and through study of the Search Manual. Students became acquainted with the problems and techniques associated with development of interest profiles including selection of terms, truncation of term fragments, development of expression for proper logical association of terms, use of links for grouping terms within an expression, and assignment of weights.

The machine-readable data base used for the student SDI experiment was Chemical Abstracts Condensates. In the first year, students conducted manual searches of an issue of Chemical Abstracts in two subject areas, one organic and the other inorganic. In the second and third years students conducted manual searches of two issues of Chemical Abstracts. After completing the manual searches, they prepared interest profiles which were used by IITRI in a search of the corresponding issues of the Condensates tapes. Output from the SDI run was returned to the students for comparison with output from their manual searches.

In many cases extremely good profiles were prepared with good relevance ratings. In other cases profiles were defective for several reasons. In all cases, after the students completed the assigned evaluation and comparison of their manual versus machine searches, they understood and were able to explain why their profiles were effective or ineffective. The time saved by the computer search was dramatic and impressed students who had had to spend considerable time in conducting the manual searches.

From the viewpoint of both instructors and students, the course accomplished its major objective, i.e., it provided a survey of traditional techniques of chemical literature, and showed the relationship of those techniques to modern search methods.

Another objective was to make the students sufficiently aware of the capabilities of computer services so that when they enter the industrial community, they will request such services. These students will be the future chemists and users of computerized chemical information systems. Hopefully, in much the same way that students who use modern analytical equipment in their university laboratories demand modern equipment in the industrial laboratories that hire them, so students familiar with automated information handling will require these services from their employers.

Miss Williams is currently discussing preparation of short courses and/or audio cassette courses based on the "Modern Techniques" course with the American Chemical Society and others.

9.4 Workbook for Modern Techniques in Chemical Information

The absence of any textbook providing adequate coverage of the modern techniques for search and retrieval of chemical information, and of the newer--principally machine-readable--sources of chemical information prompted IITRI's development of a workbook entitled Modern Techniques in Chemical Information.

The book was designed for use by chemists and does not require a background in computer technology, programming, or information science. It exposes the student to the potentials and limitations of information systems and sources and explains the storage, search, retrieval, and dissemination functions that characterize information systems.

The chapters or principal topics are: (1) "Information Systems," (2) "Indexing and Classification," (3) "Primary Information Sources in Literature," (4) "Patents," (5) "Secondary Information Sources in Literature-1: Abstracting Periodicals, Review Serials," (6) "Secondary Information Sources in Literature-2: Reference Works," (7) "Chemical Information Centers" including the computer searchable data bases and computer centers, (8) "Chemical Structures in Literature and Machine," (9) "Search Systems" including an introduction to computer components, programming languages, programming, and computer systems, (10) "Information Retrieval in a Current Awareness System."

The workbook was tested via the IIT course in 1969, 1970, 1971, and 1972. A proposal for development of a textbook based on the workbook has been submitted to NSF. After review and revision have been completed it will be published and distributed.

9.5 Technical Presentations, Publications, and Mass Mailings

The final methodology for educating potential users has been that comprising presentations at technical meetings, preparations of technical publications, and mass mailing of brief descriptions of the CSC. A listing of presentations and publications is given in Section 12 and the mass mailings are discussed in Section 10.5.1.

9.6 User Liaison

In a system that was designed to be user-oriented, frequent communication with users through various channels is extremely important. In order to maintain good rapport with users and to be sure that their profiles are functioning efficiently to provide the desired information, CSC uses many avenues of communication with users. Among them are:

- unlimited profile changes
- low-cost profile switch
- evaluation reports

- feedback cards
- continuous precision calculations
- telephone contact
- comments on profiles to suggest changes in logic weighting, and grouping of terms, or to suggest use of new data elements or new terminology
- site visits

The concern for users is of extreme importance to information centers. Information systems are designed to be used and if the clients are not satisfied with the service, they will not use it.

10. CENTER MANAGEMENT AND PROCEDURES

10.1 CSC Profile Handling Procedures

10.1.1 Receive Search Request

Search requests are received either by telephone, mail, or site visits. These requests may be made either directly by the researcher or indirectly through his representative. It is best, where possible, to discuss the search subject directly with the researcher.

10.1.1.1 Review and Interpret Search Question

The user's statement of his question is read and carefully studied. If the meaning of the question is not completely clear, the user is called to discuss his information needs. He is asked to identify pertinent search terms and synonyms, titles of pertinent papers, key authors and/or journals, etc. When the questions are received via telephone, full details are written during the call and, if possible, the requestor is asked for written confirmation. Or, he is sent a letter with the CSC interpretation of the question and/or a copy of the proposed profile for his review and comment.

10.1.1.2 Conduct Manual Search

In order to get a feel for a specific research area and to determine how this material is handled in a specific data base, a manual search is carried out. This manual search is conducted in the appropriate hard copy counterpart of the data base against which the question will run to determine useful search terms and strategy for the profile. Hard copy indexes are checked to identify additional related terms. Dictionaries, encyclopedias, etc. also assist in further defining the question and in identifying candidate terms for the profile. All worksheets prepared in development of the search strategy are kept in the profile folders.

10.1.2 Profile Handling Procedures--New Profiles

The following are the steps required in preparing a new profile for CSC:

- Review the subject of the search question and select the appropriate data base(s) against which the profile is to be run.

- Select the appropriate profile form(s). These are entitled "Computer Search Center--Search Profile-Header" (form P1), "Computer Search Center--Search Profile Terms" (form P2), and "Computer Search Center--Search Profile - Logic" (form P3). (See Figures 10-1, 10-2 and 10-3.) The forms for CA are reproduced on white paper; for EI, on yellow; and for BA, on green.

- Select candidate search terms. The selection of appropriate terms can only be done after gaining a good understanding of the search question in relation to the user's needs and in relation to the specific data base(s) against which the profile will be run.

- Check these candidate search terms for correct truncation and frequency of usage using the Truncation Guide, term lists, and/or the KLIC Index for the appropriate data base(s).

- Prepare the profile form using the profile check list. (See Figure 10-4.)

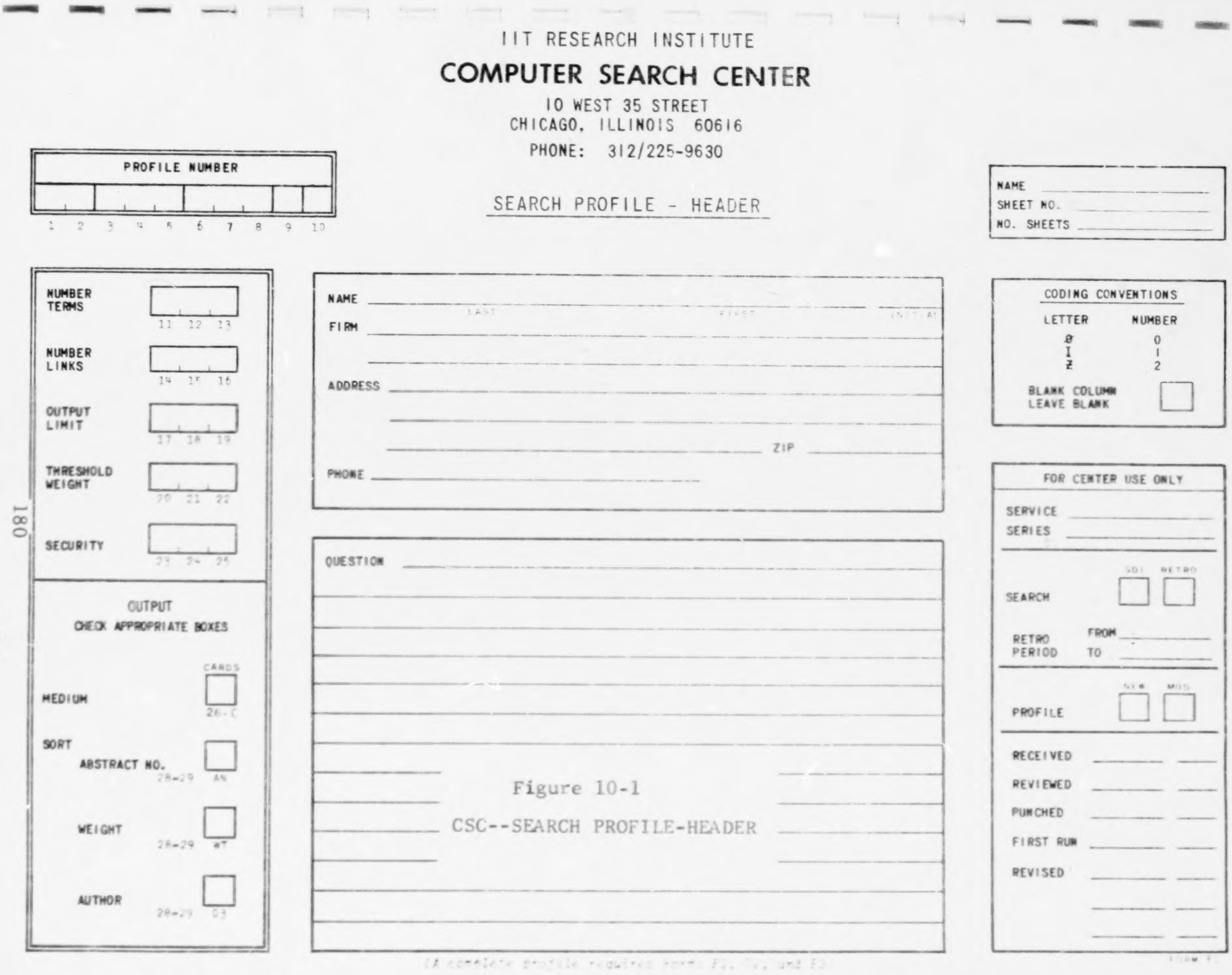
- Assign a profile number identifying the organization (corporation) from the organization code book. Profiles for a specific organization are numbered consecutively based on order of arrival. See Section 10.1.4 for details of CSC profile number designations.

- Prepare a User Record sheet for each new profile. (See Figure 10-5.) If this question comes from a new organization, a Corporate-User Record sheet must be prepared. (See Figure 10-6.)

- Prepare file cards, folders, etc., for each new profile. (A description of CSC files follows in this report).
- Xerox a copy of the completed profile. This copy is sent to the user with his first run.

- Prepare a cover letter to be sent with the output from the first run. There is a "standard" cover letter. (See Figure 10-7.) Special comments relating to a specific profile are added to this basic letter.

- The completed, checked profile is then ready to be keypunched.
- Record the profile number on the Profile Deck Modification sheet along with any appropriate comments, i.e., odd only, new profile, etc. This sheet lists all new, modified, or dropped profiles for each run. (See Figure 10-8.)



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

SEARCH PROFILE - HEADER

PROFILE NUMBER ☐☐☐☐☐☐☐☐☐☐
1 2 3 4 5 6 7 8 9 10

NAME ____________
SHEET NO. ____________
NO. SHEETS ____________

NUMBER TERMS ☐☐☐ 11 12 13

NUMBER LINKS ☐☐☐ 14 15 16

OUTPUT LIMIT ☐☐☐ 17 18 19

THRESHOLD WEIGHT ☐☐☐ 20 21 22

SECURITY ☐☐☐ 23 24 25

OUTPUT
CHECK APPROPRIATE BOXES

MEDIUM CARDS ☐ 26-C

SORT
ABSTRACT NO. ☐ 28-29 AN
WEIGHT ☐ 28-29 WT
AUTHOR ☐ 28-29 03

NAME ____________ LAST ____________ FIRST ____________ INITIAL

FIRM ____________

____

ADDRESS ____________

____

____________ ZIP ____________

PHONE ____________

QUESTION ____________

CODING CONVENTIONS

LETTER	NUMBER
Ø	0
I	1
Z	2

BLANK COLUMN LEAVE BLANK ☐

FOR CENTER USE ONLY

SERVICE ____________
SERIES ____________

SEARCH SDI ☐ RETRO ☐

RETRO PERIOD FROM ____________
TO ____________

PROFILE NEW ☐ MOD ☐

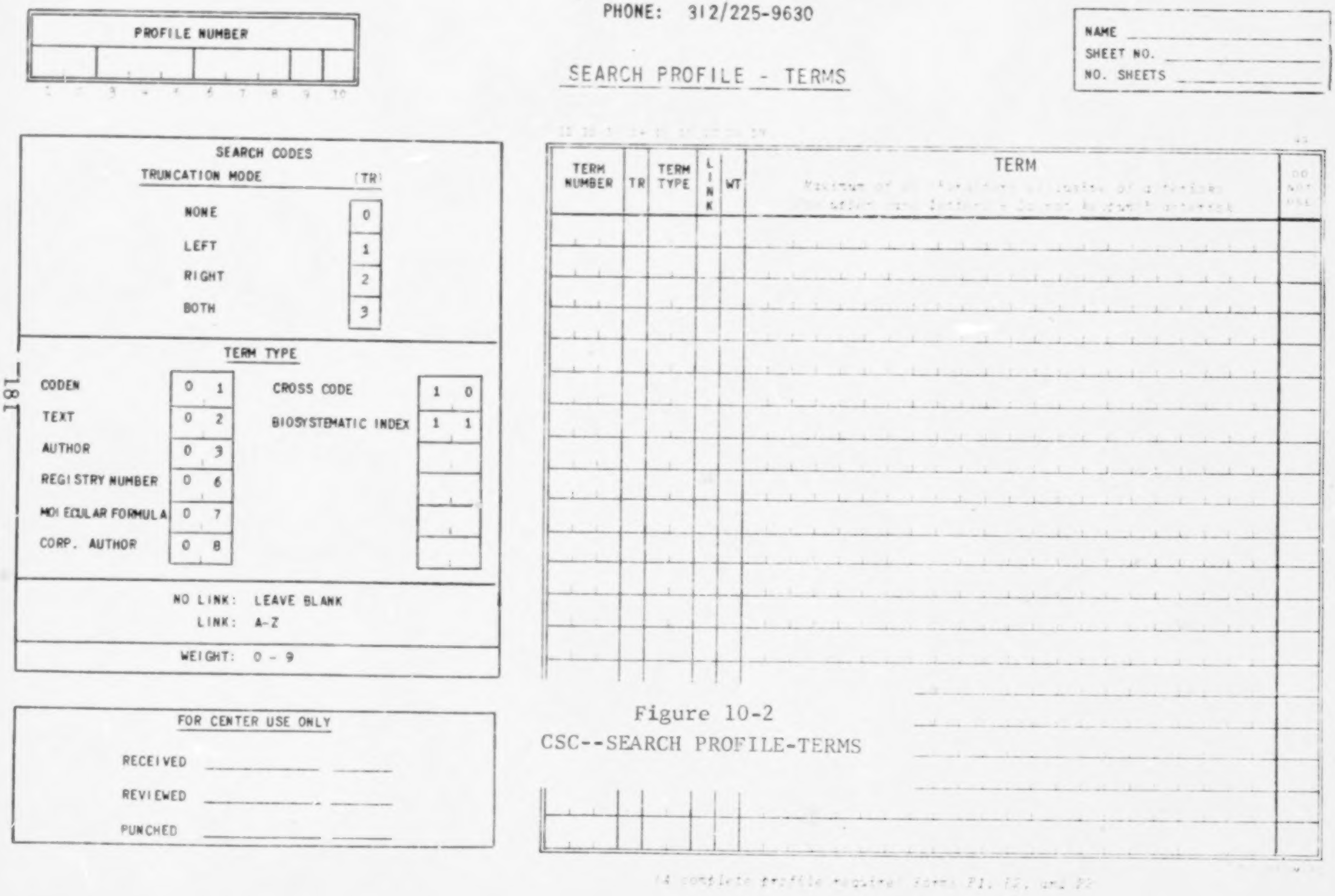
RECEIVED ____________ ____
REVIEWED ____________ ____
PUNCHED ____________ ____
FIRST RUN ____________ ____
REVISED ____________ ____

(A complete profile requires forms F1, F2, and F3)

FORM F1

Figure 10-1

CSC--SEARCH PROFILE-HEADER



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO. ILLINOIS 60616
PHONE: 312/225-9630

PROFILE NUMBER

1 2 3 4 5 6 7 8 9 10

SEARCH PROFILE - TERMS

NAME ____________
SHEET NO. ____________
NO. SHEETS ____________

SEARCH CODES

TRUNCATION MODE (TR)

NONE	0
LEFT	1
RIGHT	2
BOTH	3

TERM TYPE

CODEN	0 1	CROSS CODE	1 0
TEXT	0 2	BIOSYSTEMATIC INDEX	1 1
AUTHOR	0 3		
REGISTRY NUMBER	0 6		
MOLECULAR FORMULA	0 7		
CORP. AUTHOR	0 8		

NO LINK: LEAVE BLANK
LINK: A-Z

WEIGHT: 0 - 9

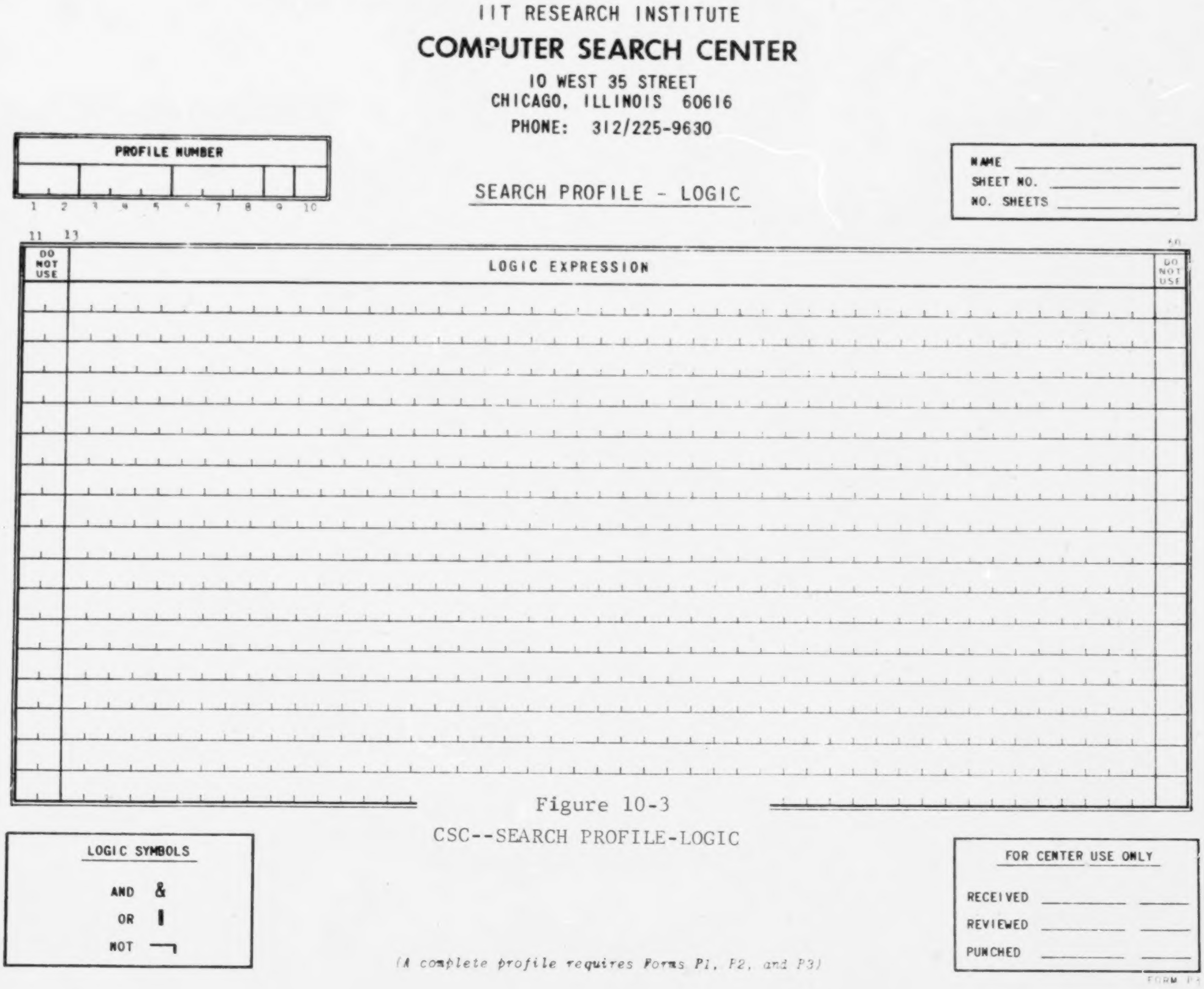
FOR CENTER USE ONLY

RECEIVED ____________ ______
REVIEWED ____________ ______
PUNCHED ____________ ______

TERM NUMBER	TR	TERM TYPE	LINK	WT	TERM [illegible]	DO NOT USE

Figure 10-2
CSC--SEARCH PROFILE-TERMS

(A complete profile requires forms P1, [illegible], and P2



IIT RESEARCH INSTITUTE

COMPUTER SEARCH CENTER

10 WEST 35 STREET
CHICAGO, ILLINOIS 60616
PHONE: 312/225-9630

SEARCH PROFILE - LOGIC

PROFILE NUMBER

1 2 3 4 5 6 7 8 9 10

NAME ____________
SHEET NO. ____________
NO. SHEETS ____________

11 13 60

DO NOT USE	LOGIC EXPRESSION	DO NOT USE

LOGIC SYMBOLS

AND &
OR |
NOT ¬

FOR CENTER USE ONLY

RECEIVED ____________
REVIEWED ____________
PUNCHED ____________

(A complete profile requires Forms P1, P2, and P3)

FORM P4

Figure 10-3

CSC--SEARCH PROFILE-LOGIC

Check each of the following as considered

____Profile form correct for data base to be searched, CA - white
EI - yellow
BA - green.

____Profile number coded correctly according to data base and user-type.

____Profile number recorded on each page.

____User name and address correct and complete.

____User phone number recorded and complete.

____User name recorded on each page of profile.

____Each page of profile form numbered.

____Statement of search question as detailed as possible. All available information recorded.

____CA search coverage (even, odd, both) recorded.

____# of terms recorded corresponds to # of terms listed.

____# of links recorded corresponds to # of links listed.

____Output limit recorded.

____Threshold weight recorded. Does this correctly represent the question? If weights assigned, do all terms have weight recorded? Do the "not" terms have zero weight?

____Medium and Sort recorded.

____Letters, "Ø", "I" and "Ƶ", and numbers "0", "1" and "2" correctly written.

____Term-types are correctly recorded.

____Terms correctly spelled.

____Term-truncation and frequencies checked.

____Truncation modes are correctly recorded.

____Terms are correctly numbered.

____Each link contains 2 or more terms.

____All terms and links are accounted for in the logic.

____Logic statement correctly expresses search question.

____Logic statement is clearly and correctly printed with brackets in place.

____Modifications are recorded, dated and initialled.

5/4/72 - PAL

Figure 10-4

PROFILE PREPARATION CHECK LIST

User Name ______________________________ Phone ______________

Address ______________________________

____

____

____

Profile Identification					Issues			comments	rec'd	first run	dropped
svc	company code	user no.	no.	mod	O	E	B				

Figure 10-5

USER RECORD SHEET

CORPORATE-USER RECORD

MAILING ADDRESS: ______________________________

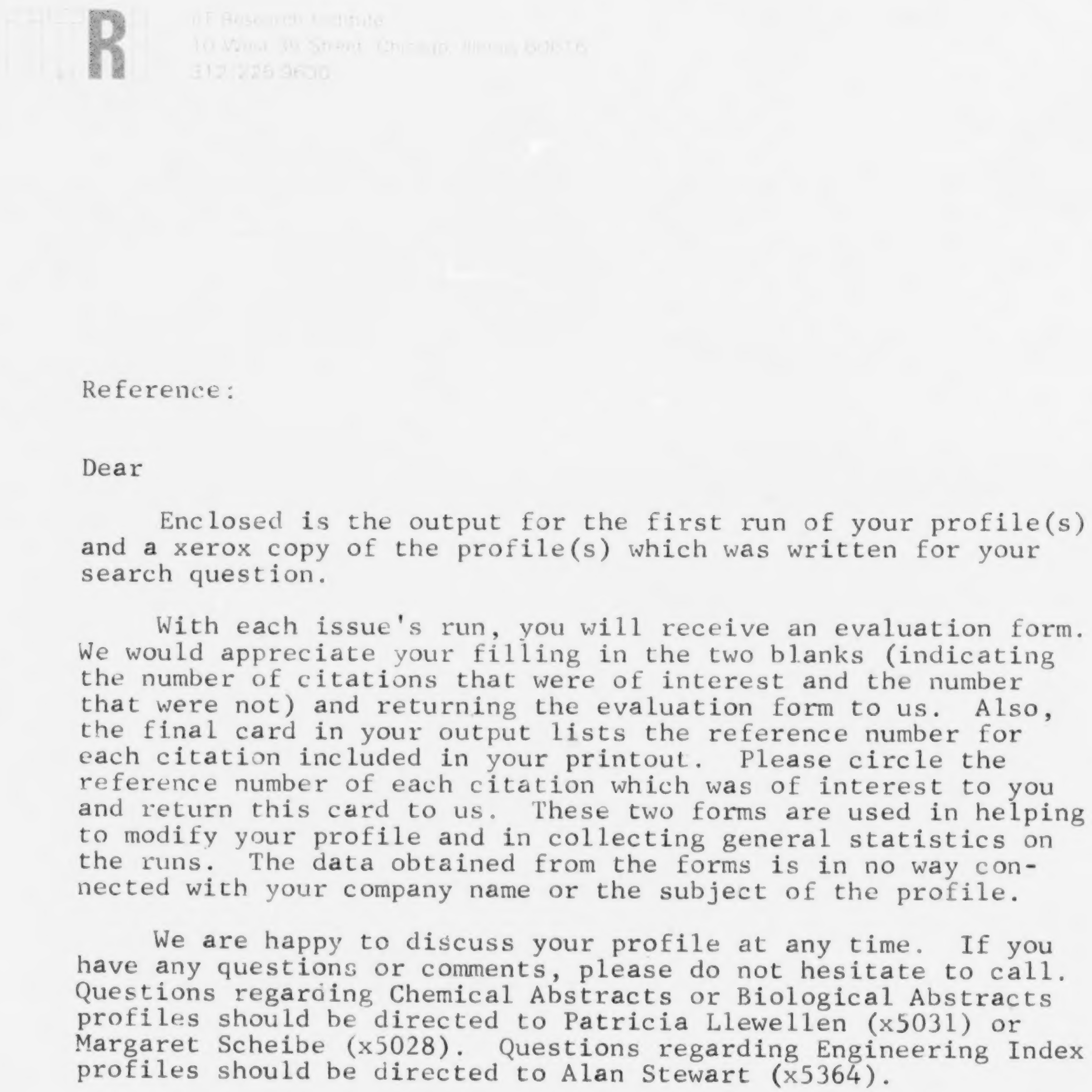
Attn: ______________________________

Phone: ______________________________

user code	name	number
001		1 2 3 4 5 6 7 8 9
002		1 2 3 4 5 6 7 8 9
003		1 2 3 4 5 6 7 8 9
004		1 2 3 4 5 6 7 8 9
005		1 2 3 4 5 6 7 8 9
006		1 2 3 4 5 6 7 8 9
007		1 2 3 4 5 6 7 8 9
008		1 2 3 4 5 6 7 8 9
009		1 2 3 4 5 6 7 8 9
010		1 2 3 4 5 6 7 8 9
011		1 2 3 4 5 6 7 8 9
012		1 2 3 4 5 6 7 8 9
013		1 2 3 4 5 6 7 8 9
014		1 2 3 4 5 6 7 8 9
015		1 2 3 4 5 6 7 8 9
016		1 2 3 4 5 6 7 8 9
017		1 2 3 4 5 6 7 8 9
018		1 2 3 4 5 6 7 8 9
019		1 2 3 4 5 6 7 8 9
020		1 2 3 4 5 6 7 8 9

Figure 10-6

CORPORATE-USER RECORD SHEET



IIT Research Institute
10 West 35 Street, Chicago, Illinois 60616
312/225-9630

Reference:

Dear

Enclosed is the output for the first run of your profile(s) and a xerox copy of the profile(s) which was written for your search question.

With each issue's run, you will receive an evaluation form. We would appreciate your filling in the two blanks (indicating the number of citations that were of interest and the number that were not) and returning the evaluation form to us. Also, the final card in your output lists the reference number for each citation included in your printout. Please circle the reference number of each citation which was of interest to you and return this card to us. These two forms are used in helping to modify your profile and in collecting general statistics on the runs. The data obtained from the forms is in no way connected with your company name or the subject of the profile.

We are happy to discuss your profile at any time. If you have any questions or comments, please do not hesitate to call. Questions regarding Chemical Abstracts or Biological Abstracts profiles should be directed to Patricia Llewellen (x5031) or Margaret Scheibe (x5028). Questions regarding Engineering Index profiles should be directed to Alan Stewart (x5364).

Figure 10-7

CSC STANDARD "COVER" LETTER

10.1.3 Profile Handling Procedures--Modified Profiles

A copy of the current version of the profile must be maintained in the file. Xerox this profile (N.B., return the original to the file) and use the copy as a working copy during modification.

- Attach a complete new form Pl to the profile form. For CA this is white; for EI, yellow; and for BA, green.
- Make all necessary changes on the Xerox copy of the profile, date and initial all changes on form Pl, and record the reasons for the change(s). Where changes are extensive, prepare a complete new set of profile forms.
- Assign a modification number. This involves a change in the tenth character of the profile number, e.g., A→B→C→D, etc.
- A Xeroxed copy of this modified profile is made. Send this to the user with the first output from this modification.
- Make necessary changes in the profile records, i.e., change user name, output limit, output frequency, etc.

10.1.4 Designation of CSC Profile Numbers

10.1.4.1 CSC Profile Number

A CSC profile number consists of ten alphanumeric characters in the following form: AN-ANN-NNN-NA (A indicates a letter; N, a number).

Character	1	indicates the data base, e.g., B indicates BA, C indicates CA, and E indicates EI.
Character	2	indicates odd (1), even (2), or both (3) issues of CA. (1) used for BA & EI.
Character	3	indicates user-type classification.
Characters	4-5	indicate the corporate number (the user).
Characters	6-7-8	indicate the user within a corporation.
Character	9	indicates the profile (of a user).
Character	10	indicates the modification version.

Date:

PROFILE DECK MODIFICATIONS

Service:
Vol/Iss:

PROFILE NO.	NEW	MOD	DROP	DATE OF MINIPUP	COMMENTS

Figure 10-8
PROFILE DECK MODIFICATION SHEET

10.1.4.2 User-Type Classification

General classifications that indicate type of user are as follows:

A-F	Academic
G-K	Independent Research Organization
L-R	Industrial
S	Workshop
W&Y	Government
X	Experimental and Standard Profiles

10.1.5 Keypunch Profile

All profiles--new or modified--are entered into the system within five working days of receipt by CSC. Keypunching is scheduled to meet this objective. All keypunching is proofread twice, first by the keypuncher and second by someone other than the keypuncher.

10.1.6 Enter Profile in Input Data Deck

The profile keypunch cards and the Profile Deck Modifications sheet are correlated. These records are later checked against DKEDIT and MINIPUP to assure all new or modified profiles are accounted for.

10.1.7 Check Output and Prepare Mailing

Before the output is packaged, output for new profiles and revised profiles must be carefully checked. The retrieved citations are reviewed for technical value to the search question. Consideration is given to the value of material in the data base, non-pertinent citations due to faulty logic, misinterpretation of a concept, etc. Search terms are rechecked for spelling, truncation, term type designations, and search logic in relation to the particular citations retrieved. If a serious error or problem has occurred, the user is called to discuss this problem and to discuss the procedures required to correct it before the next run. With each new profile, the cover letter (see Figure 10-7) is sent to the user requesting his return of the evaluation report sheet (see Figure 8-1) and trailer card (see Figure 10-9) and explaining how they are

```
C1L490102A    CA CONDENSATES HITS FOR VOL. 76, NO. 15     APRIL 15, 1972

          081072  084003  085494
          081685  084004
          081748  084439
          082416  084503
          082529  084579
          082964  084580
          082988  084631
          082989  084685
          082992  084687
          082993  084693
          082996  084698
          082998  084726
          083063  084757
          083136  084777
          083159  084780
          083162  084785
          083164  084786
          083173  084789
          083201  084807
          083232  084813
          083249  084817
          083345  084868
          084002  085234
```

Figure 10-9
CSC TRAILER CARD

filled out. A copy of the profile is sent with the first run from each new or modified profile.

After the output is received and reviewed it is packaged and labeled for delivery.

10.1.8 Monitoring Profiles

10.1.8.1 Monitoring New Profiles

All new profiles are carefully monitored for the first four or five runs. User evaluation forms are checked for specific comments on the output as well as for identification of pertinent and non-pertinent citations. If these evaluations are not returned promptly the user should be called to discuss the output. Typical questions to ask him are: Has the output been relevant? Have there been specific problem areas in all the output sent so far (i.e., have all non-relevant references come from a given section of CA)? Does the user know of "missed" citations? Discuss the search question again. Have the search results pointed out areas of interest or non-interest that the user had not considered before? Determine what changes are necessary to improve the usefulness of the output. Review user requests to add, delete or change terms and/or logic carefully. If there appear to be problems implementing the request, call the user to discuss his new information needs.

10.1.8.2 Monitoring Existing Profiles

After a profile has been stabilized, output is checked every four to six runs to be sure no error or change in data base format is affecting the search results. The user is called every two to three months to 1) check on the performance of the profile, 2) apprise him of any new searchable data elements, and 3) determine if his information needs are changing due to changes in his research interests. If the profile needs modifications, these should be discussed and implemented. User requests for changes should be carefully reviewed to determine their effect on the profile output.

10.1.9 Dropping of Profiles

A user may request that a specific profile be dropped. The appropriate file changes and deletions must be made. These include:

- Prepare two DROP cards--one for the term and one for the logic section of the deck. A DROP card has the profile number in columns 1-10 and the word DROP in columns 11-14.
- Complete a CSC Profile DROP Checklist for each Corporate User Record sheet. (see Figure 10-10)
- Complete the User Record sheet. (see Figure 10-5)
- File Profile folder in Dropped file drawer.
- Record "DROP" status on User subscription card, on Profile Hit Record Sheet(s), and in Profile History Book.

10.2 Center Files and User Records

10.2.1 User Record File

Individual user records are maintained on 8-1/2" x 11" sheets. Each sheet includes the user's name, company affiliation, mailing address, telephone number, and individual profile number. Profile modification and status are recorded. These sheets are filed alphanumerically by corporate and user code.

10.2.2 Corporate-User Record File

This file is made up of 8-1/2" x 11" sheets which identify company name, address, corporate code, company contact(s), and telephone number(s). They are arranged alphanumerically by corporate code in a Corporate-User Record File book. The Corporate User Record sheet also indicates the name of each user within the company and the number of profiles he has running in the system. (see Figure 10-6)

10.2.3 CSC User Profile History Book (Restricted Data)

This book contains detailed data on profiles for each CSC user. Information in this book includes individual user name, corporate name and telephone number, status of profile, and dates. This information is arranged alphabetically by corporate user. This material is updated daily.

______ Corporate Code

CSC Profile DROP Checklist

______ Date

______ Initial

Profile number(s) ______ ______

______ ______

______ ______

______ ______

______ ______

Reason(s) for dropping: ___ end of free-trial, not purchased

___ end of subscription, not repurchased

Give specific reason(s) for termination of services:

______

______

______

______

______

Procedural check:

___ "Drop" status recorded on Corporate User Sheet

___ User Record Sheet completed, pulled and refiled in profile folder

___ Profile folder refiled in appropriate "dropped" file drawer

___ User subscription cards pulled and refiled

___ "Drop" status and date recorded on Profile Hit Record Sheet(s)

___ "Drop" cards keypunched and submitted to appropriate deck(s)

___ "Drop" status and date recorded in Profile History Book.

Figure 10-10
DROP CHECKLIST

10.2.4 CSC Profile Folder Files

Active Profile Folders files are arranged alphanumerically by profile number for each data base. There is a folder in the file for each active profile number. This folder contains a copy of each profile modification. A master folder of company correspondence and contact information is filed immediately in front of each company's set of individual user profiles.

Inactive (dropped) Profile Folder files are filed alphanumerically by profile number.

10.2.5 Profile Correspondence File

Correspndence related to specific profile activities, e.g., term additions, deletions, modifications, etc., is filed in a folder directly in front of each company's set of profiles.

10.2.6 Telephone Number File

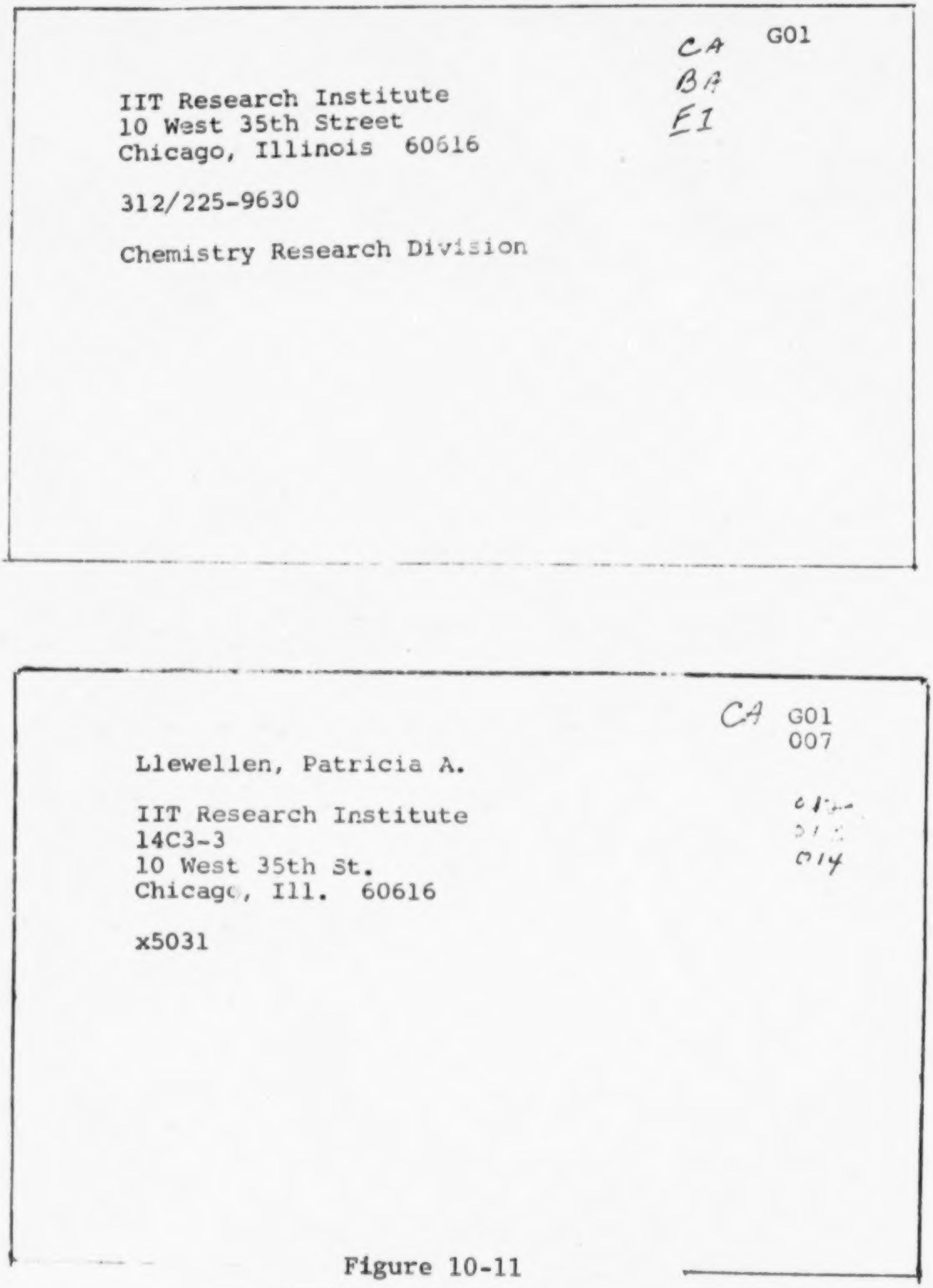
The telephone number file is maintained on 4" x 6" cards. Phone numbers are referenced in two ways. One half of the file is arranged by company name. The other half is arranged by user name. This file is used: for easy access to company and/or individual user telephone number or for rapid identification of corporate code. Many user contacts (by letter or telephone) are made without identifying profile numbers. Profile file locations can be identified rapidly using this file. (See Figure 10-11.)

10.2.7 Profile Hit Evaluation File

This file is made up of 11" x 17" fold-out sheets on which are recorded weekly (CA) profile hit statistics. The number of hits received per search and user evaluation results are recorded for each profile. These sheets are arranged alphanumerically according to code. (See Figure 10-12.)

10.2.8 Evaluation Report File

On these 8½" x 11" sheets are recorded profile number, user name,



CA
BA
EI

G01

IIT Research Institute
10 West 35th Street
Chicago, Illinois 60616

312/225-9630

Chemistry Research Division

CA

G01
007

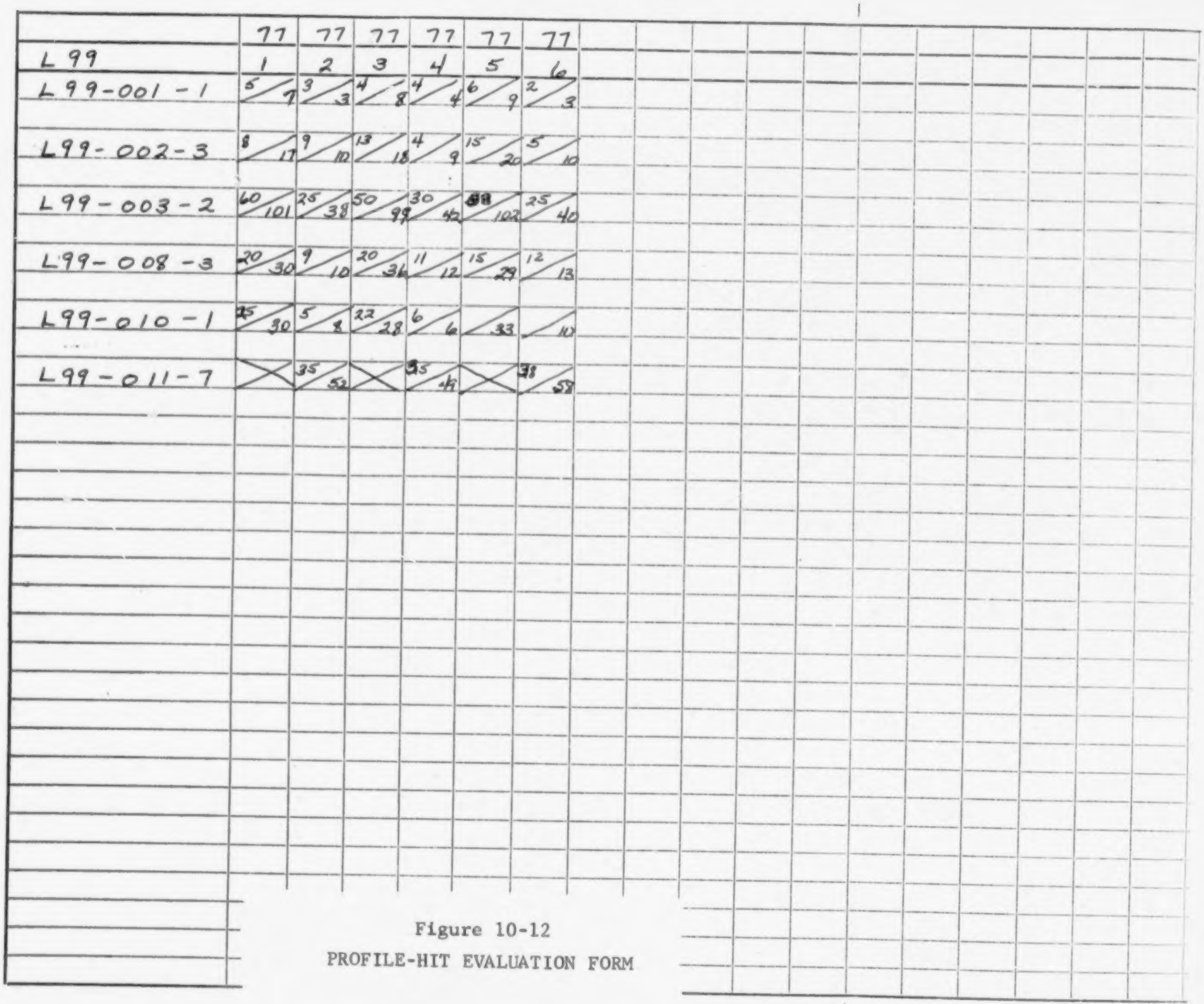
Llewellen, Patricia A.

IIT Research Institute
14C3-3
10 West 35th St.
Chicago, Ill. 60616

x5031

01[illegible]
[illegible]
014

Figure 10-11
TELEPHONE NUMBER FILE



	77	77	77	77	77	77
L 99	1	2	3	4	5	6
L99-001-1	5/7	3/3	4/8	4/4	6/9	2/3
L99-002-3	8/17	9/10	13/18	4/9	15/20	5/10
L99-003-2	60/101	25/38	50/99	30/42	[illegible]/102	25/40
L99-008-3	20/30	9/10	20/36	11/12	15/29	12/13
L99-010-1	25/30	5/8	22/28	6/6	/33	/10
L99-011-7	X	35/52	X	35/45	X	38/58

Figure 10-12
PROFILE-HIT EVALUATION FORM

number of citations (hits) sent, and the user's evaluation and comments concerning these citations. They are filed alphanumerically by profile number in groups according to data base volume. These reports are reviewed weekly to assist in maintenance and modifications of each user's profiles.

Statistics relating to profile hit relevance are prepared from these sheets. Evaluation data from these sheets are recorded in the Profile Hit Evaluation Record file. (See Figure 10-12.)

10.2.9 Abstract Number Card File

These cards are the trailer cards from each CSC profile output. Recorded on each card is the user evaluation of the hits for his profile. These evlauations are reviewed to assist in profile maintanance and updating procedures. These 5" x 8" cards are filed by profile number in data base volume groups.

10.2.10 Profile Deck Modifications File

These 8" x 11" sheets are completed weekly as new profiles or profile modifications are generated. They accompany the profile keypunch cards and are checked at time of DKEDIT to assure all new modified or dropped profiles are in the system. These sheets are filed according to data base by volume and issue number. (See Figure 10-8.)

10.2.11 CSC Profile DROP Checklist

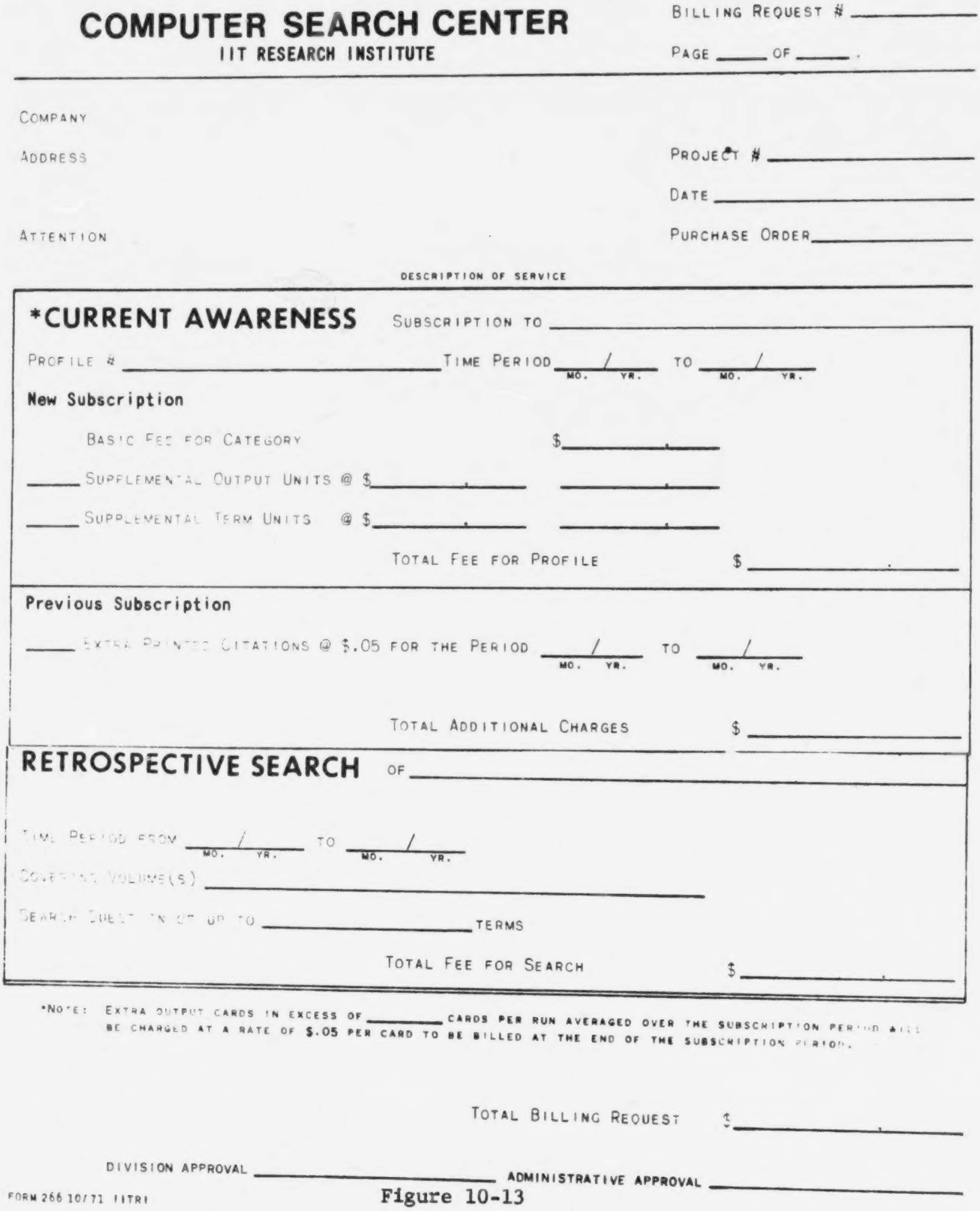
These 8½" x 11" sheets are prepared for each profile to be dropped from the system. After the various "drop" procedures are completed, this sheet is filed in a folder in corporate code order in the drawer with the dropped profiles.

10.2.12 CSC Billing Forms

Billing Forms are prepared for each profile. They are maintained in alphabetic order by company and are cross referenced to the billing number. (See Figure 10-13.)

10.2.13 Subscription Card File

This 5" x 8" card file, organized alphabetically by company



COMPUTER SEARCH CENTER
IIT RESEARCH INSTITUTE

BILLING REQUEST # ________
PAGE ____ OF ____.

COMPANY
ADDRESS
PROJECT # ________
DATE ________
ATTENTION
PURCHASE ORDER ________

DESCRIPTION OF SERVICE

***CURRENT AWARENESS** SUBSCRIPTION TO ________

PROFILE # ________ TIME PERIOD ___/___ (MO. YR.) TO ___/___ (MO. YR.)

New Subscription

BASIC FEE FOR CATEGORY $____.____

____ SUPPLEMENTAL OUTPUT UNITS @ $____.____ ____.____

____ SUPPLEMENTAL TERM UNITS @ $____.____ ____.____

TOTAL FEE FOR PROFILE $____.____

Previous Subscription

____ EXTRA PRINTED CITATIONS @ $.05 FOR THE PERIOD ___/___ (MO. YR.) TO ___/___ (MO. YR.)

TOTAL ADDITIONAL CHARGES $________

RETROSPECTIVE SEARCH OF ________

TIME PERIOD FROM ___/___ (MO. YR.) TO ___/___ (MO. YR.)

COVERING VOLUME(S) ________

SEARCH QUESTION OF UP TO ________ TERMS

TOTAL FEE FOR SEARCH $____.____

*NOTE: EXTRA OUTPUT CARDS IN EXCESS OF ______ CARDS PER RUN AVERAGED OVER THE SUBSCRIPTION PERIOD WILL BE CHARGED AT A RATE OF $.05 PER CARD TO BE BILLED AT THE END OF THE SUBSCRIPTION PERIOD.

TOTAL BILLING REQUEST $____.____

DIVISION APPROVAL ________ ADMINISTRATIVE APPROVAL ________

FORM 266 10/71 IITRI

Figure 10-13
BILLING FORM

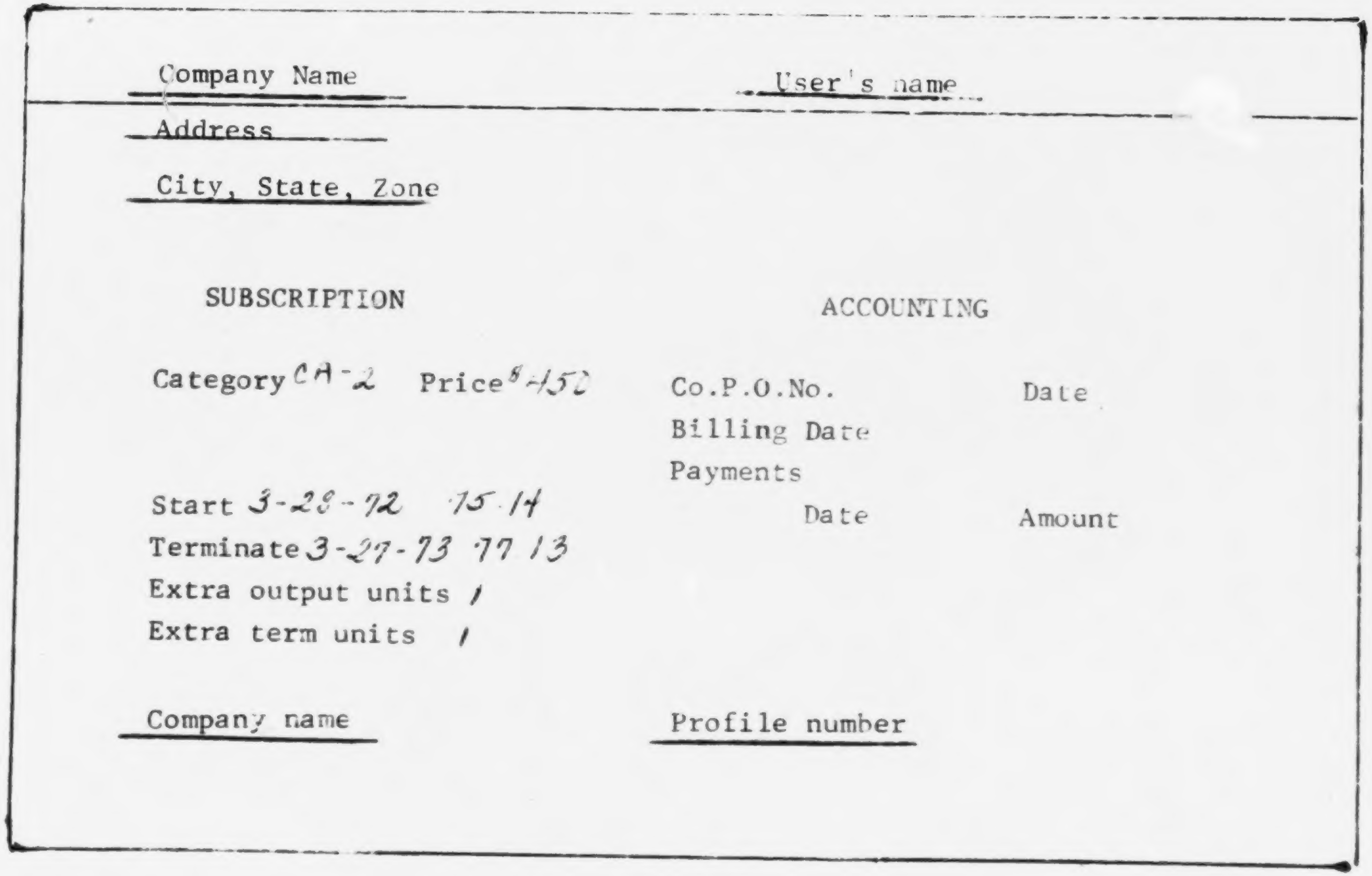
name, gives subscription and billing information including coverage, starting and terminating dates, etc. (See Figure 10-14.)

10.2.14 Profile Subject Index

Profile subject categories are entered on 5"x 8" white cards along with their related profile number(s). Term cross-indexing has been done for major subject categories. This file, arranged alphabetically by subject, is used by the CSC staff in profile preparation to locate profile questions with similar subject coverage. (See Figure 10-15.)

10.2.15 Form Masters and Supplies

Master (reproducible) copies of all CSC forms are maintained in the CSC office. Supplies of these forms are kept in this office. When these supplies become diminished, the form is reviewed for possible updating or other revisions. A modified form is prepared when necessary, copies are made, and the "new" master is set aside in the file.



Company Name

User's name

Address

City, State, Zone

SUBSCRIPTION

Category CA-2 Price $450

Start 3-28-72 75.14
Terminate 3-27-73 77.13
Extra output units 1
Extra term units 1

Company name

ACCOUNTING

Co.P.O.No. Date
Billing Date
Payments
Date Amount

Profile number

Figure 10-14
SUBSCRIPTION CARD

POLLUTION

B 23-006-1
E 14-009-3
L 19-997-4
C 60-002-8
B 24-007-1
C 45-007-2
B 56-009-4
C 40-005-3
L 79-008=2
L 69-016-7
B 52-005-5
C 70-001-7
C 54 005-2
C 32-009-7
B 23-011-3
B 56-019-2
L 45-007-3

Figure 10-15

PROFILE SUBJECT INDEX

10.3 Tape Quality and Handling

10.3.1 Tape Quality--Physical Characteristics

The major difficulty in this respect is that of physical damage to the tapes. This results in an inability to read the tape into core. Other than obvious physical damage such as being run over by a truck (this has happened--tire marks were visible), we have received tapes that were dirty and/or written with a tape drive that had mechanical defects which caused misalignment of bits (skew errors). Dirty tapes (a thin film is sufficient to damage tape) can sometimes be rescued by operator intervention and cleaning on the tape drive. This is a poor solution, since it causes the operator to interrupt processing of all jobs. About six of the 52 CA tapes issued in 1971 were damaged in this way. Three were corrected by cleaning, and three were replaced. We had two such tapes from EI and none from BA. In 1972 we have received two damaged tapes from EI.

10.3.2 Tape Quality--Readable, Mis-recorded Information

The second area in which we have experienced problems is that of wrong information on tapes. This includes machine-readable labels that do not correspond to paper labels. At present data base paper labels, as they are sent from suppliers, are inadequate. As a minimum, a label should denote:

- tracks
- recording density
- reel number
- number of files on tape
- record and block size
- supplier name and address
- creation date & job number under which created
- dataset name of each file

The other errors of this type refer to data that are coded incorrectly, for example, directory entries that are wrong. CSC programs for conversion skip non-acceptable data, and if enough portions of a citation are garbled the entire citation is skipped. In 1971 CA Condensates had about one bad record per 7000. It is now about 1 in 20,000. Since EI and BA tapes are constructed differently, we process the errors as given, and so cannot tell how many have occurred for EI.

10.3.3 Tape Quality--Wrong Information

The final category includes misspellings and similar errors for which we do not make checks. However, from such things as KLIC indexes, we do know that these errors occur. For example, of the 351 EI Card-a-Lert codes found on the 1971 tapes, 131 were spurious. Some 10% of the words in the alphabetical term list for EI are misspelled. BA and CA have less than 1% of this type of error, based on observation, not actual count.

10.3.4 Tape Handling Procedures

We have established the following procedures for detecting these conditions and converting data bases to IITRI-format. When tapes are received, they are logged in and sent to the data center. Here the paper label is checked, the appropriate format conversion program is chosen, and the JCL is prepared. The program is then run. If the conversion program runs properly, the output is checked for "bad" records. If these are few, the converted tape is used for the production run and the original is copied for backup storage. If the number was greater than fifty, we obtain a new tape from the supplier, returning the original and as much information as possible to inform the supplier of the errors.

If the format conversion program does not run properly, there are four possible causes.

(1) If the wrong JCL is employed, the entire tape would be unreadable. In this case we correct JCL and run again.

(2) If the machine-readable label on the tape contained an error, the entire tape would be unreadable. In this case we dump the label and first few records. If the dump indicates a blank tape, we obtain a replacement. If it indicates a wrong density or mis-labelling, we change our JCL to conform to the actual data

and rerun the conversion program. However, we notify the supplier of errors.

(3) If the tape contains dirt, oxide film, a crease, crinkle or other physical defect, the format conversion program will fail during processing. At times the tape is merely dirty and can be cleaned. This will be attempted. If salvage is not possible, we will obtain a replacement, and provide the supplier with documentation as to type and position of error.

(4) A change in data base format may cause the format program to fail completely or to run but produce an incorrect conversion. Determination of this kind of error requires a dump and analysis of the incoming tape, and the only remedy is to modify the conversion program to take into account the format change.

10.4 Production Statistics and Cost

Detailed statistics have been collected on operations and costs associated with the Computer Search Center. Direct search costs are relatively easy to obtain inasmuch as these costs are derived from computer processing. Production statistics and costs for searches of CA, BA, and EI are given in this section.

10.4.1 Computer Time per Program

The overall programming system is made up of five basic programs (Section 5). CSC monitors the system by continuously checking the amount of time (cost) and the relative percent time each of the individual programs expends in carrying out production runs. Data presented here have been normalized for purposes of comparison. Tables giving the percent of computer time for the four program functions: Format Conversion (data base input preparation), Input (profile input preparation), Search, and Output (preparation of output) are given in Tables 10-1 through 10-13. The fifth program, Statistics generation, uses so little computer time that it was omitted from these tabulations. Data showing the same relationships have been graphed and are presented in Figures 10-16 through 10-24.

The percentage of computer time is a relative number--as the percentage of one program decreases the percentages of the other programs increase. However, absolute cost of the entire system decreases as the cost for any individual program decreases. Examination of the percentages helps us determine which portions of the system (programs or modules) we want to work on to further cut costs. Table 10-1 gives the average percent run times and ranges of percent run time for processing CA Volume 76.

Program Function	Program or Module	Average % Time	Range % Time
Data base input preparation	FORCON	10.90	9.0-14.0
Profile input preparation	DKEDIT-MINIPUP INPUTR	1.56 2.05	1.0- 4.0 2.0- 4.0
Search (term match and logic evaluation	SEARCH	75.06	68.0-75.0
Output preparation	OCP PRINT	6.38 4.21	5.0- 7.0 3.0- 5.0
Statistics generation	STIXA	.17	0.1- 0.5
Private Libraries extraction	PLSXT	.57	0.5- 1.0

Table 10-1

PERCENT AND RANGE PERCENT COMPUTER TIME PER PROGRAM

Table 10-2 displays for comparison the average percent computer time per program for CA Volumes 71 and 76.

Program Function	Program or Module	Average % Time Vol. 71	Average % Time Vol. 76
Data base input preparation	FORCON	16.71	10.00
Profile input preparation	DKEDIT-MINIPUP INPUTR	1.43	3.61
Search	SEARCH	77.54	75.06
Output preparation	OCP-PRINT	4.32	11.16*
Statistics generation	STIXA	-	.17

*includes private libraries extraction--0.57

Table 10-2

AVERAGE PERCENT COMPUTER TIME PER PROGRAM

CHEMICAL ABSTRACTS CONDENSATES VOLUME 71

Issue	Format Conversion		Input		Search		Output	
	Time*	%	Time*	%	Time*	%	Time*	%
(Data for issue Nos. 1 - 8 does not exist)								
9	5:48	2.03	10:11	3.56	4:12:27	88.30	17:26	6.10
10	11:44	2.56	13.17	2.91	6:21:13	83.50	2:48	11.04
11	4:55	1.06	9:22	2.02	6:46:28	87.64	2:23	9.28
12	5:46	1.11	8:28	1.63	7:31:13	86.84	54:09	10.42
13	14:00	3.07	21:08	4.58	6:38:22	85.69	30:44	6.66
14	16:05	3.41	9:15	1.96	6:39:21	84.68	46:55	9.95
15	18:43	7.91	18:31	7.82	3:11:21	80.84	8:07	3.43
16	17:13	9.39	5:24	2.95	2:28:16	80.89	12:23	6.76
17	20:43	10.13	9.38	4.71	2:22:54	69.85	31:19	15.31
18	Test	-	-	-	-	-	-	-
19	13:36	7.14	8:29	4.45	1:46:33	55.93	1:01:51	32.47
20	17:28	8.51	6:19	3.08	2:22:27	69.42	38:58	18.99
21	11:45	5.73	7:55	3.86	2:14:53	65.73	50:39	24.68
22	15:11	5.63	6:19	2.34	2:59:23	66.51	1:08:50	25.52
23	8:28	4.15	11:19	5.96	2:41:03	84.81	10:45	5.66
(Data for issue Nos. 24 - 26 do not exist)								
Odd	12:01	5.15	11:09	4.62	3:43:53	77.35	26:39	12.95
Even	13:54	5.10	8:08	2.48	4:43:39	8.64	37:21	13.78
Total	12:49	5.13	9:52	3.70	4:19:26	77.90	31:14	13.31

Time* = Normalized time = $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

Table 10-3
PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	13:35	7.66	11:51	6.68	2:20:07	79.03	11:45	6.63
2	19:54	10.35	10:54	5.67	2:31:24	78.73	10:06	5.25
3	13:39	7.65	12:46	7.15	2:13:38	74.86	18:27	10.34
4	18:58	9.26	11:43	5.72	2:43:38	79.86	10:34	5.16
5	13:44	8.62	12:27	7.82	2:00:22	75.51	12:40	7.95
6	27:12	8.88	13:41	4.47	4:05:17	80.08	20:07	6.57
7	14:40	7.85	13:06	7.01	2:22:34	76.28	16:34	8.86
8	24:04	7.88	14:14	4.66	4:09:09	81.56	17:57	5.88
9	18:15	7.49	12:28	5.12	3:09:31	77.80	23:22	9.59
10	20:37	6.80	12:55	4.26	4:12:53	83.38	16:55	5.58
11	17:26	7.14	13:19	5.45	3:12:23	78.78	21:04	8.63
12	23:19	6.78	14:43	4.28	4:46:05	83.21	19:42	5.73
13	10:25	4.58	14:03	6.18	3:02:13	80.13	20:43	9.11
14	24:28	6.28	15:52	4.07	5:27:02	83.92	22:20	5.73
15	16:18	6.56	14:23	5.79	3:16:19	79.03	21:25	8.62
16	25:52	6.55	2:17	0.58°	5:45:58	87.63	20:41	5.24
17	18:45	7.93	1.48	0.76	3:14:56	82.46	20:55	8.85
18	21:27	8.10	1:50	0.19	3:49:19	86.57	12:17	4.64
19	18:43	8.36	1:43	0.77	3:04:58	82.65	18:24	8.22
20	24:35	9.20	2:29	0.93	3:55:51	84.12	15:22	5.75
21	15:10	9.56	1:58	1.24	1:59:24	75.24	22:09	13.96
22	18:26	14.67	1:55	1.52	1:32:48	73.82°	12:33	9.99
23	23:05	21.44	22:17	2.12°	1:04:59	60.34	17:20	16.10
24	27:38	20.94	2:15	1.71	1:39:07	75.09	2:59	2.26°
25	14:08	14.73	:33	2.65	1:15:03	78.13	4:19	4.49
26	29:49	21.89	2:10	1.59	1:41:02	74.18	3:11	2.34
Odd	15:59	9.20	8:49	4.52	2:28:58	76.95	17:37	8.81
Even	23:34	10.58	8:14	3.09	3:32:58	80.94	14:13	7.11
Total	19:46	9.89	8:31	3.80	3:00:58	78.94	15:55	7.96

Time* = Normalized time - $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

o = Major Modifications

Table 10-4

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	9:14	11.45	2:26	3.02	1:05:12	80.79	3:50	4.74
2	28:27	17.79	2:49	1.76	2:04:19	77.75	4:19	2.70
3	21:42	15.79	2:43	1.98	1:47:57	78.57	5:02	3.66
4	15:40	11.35	2:29	1.80	1:55:49	83.93	4:02	2.92
5	21:43	17.65	2:23	1.94	1:34:28	76.80	4:26	3.61
6	22:31	21.57	2:12	2.10	1:16:37	73.38	3:05	2.95
7	17:27	16.38	2:21	2.21	1:12:08	77.12	4:02	3.79
8	25:49	25.38	2:09	2.11	1:09:58	68.79	3:47	3.72
9	19:39	20.27	2:17	2.36	1:10:30	72.76	4:28	4.61
10	24:50	23.32	2:08	2.00	1:15:43	71.09	3:49	3.59
11	15:25	15.11	2:23	2.33	1:19:06	77.55	5:13	5.11
12	19:56	19.04	2:18	2.19	1:18:33	75.02	3:56	3.75
13	17:49	13.75	2:22	1.82	1:44:26	80.56	4:59	3.85
14	22:36	21.64	2:02	1.94	1:15:44	72.54	4:03	3.88
15	20:08	19.07	2:17	2.17	1:18:02	73.90	5:08	4.86
16	23:19	20.14	2:10	1.87	1:25:06	73.49	5:13	4.50
17	15:52	13.63	2:15	1.94	1:32:36	79.56	5:40	4.87
18	25:28	22.46	2:01	1.77	1:20:19	70.82	5:37	4.95
19	14:23	11.72	2:22	1.93	1:39:34	81.14	6:24	5.21
20	26:25	21.69	2:09	1.76	1:27:00	71.43	6:14	5.12
21	21:48	15.94	2:18	1.68	1:45:57	77.45	6:45	4.93
22°	21:45	18.69	2:07	1.82	1:25:37	74.56	6:16	5.38
23°°	17:29	13.43	2:25	1.86	1:49:59	79.09	6:42	5.15
24	23:59	20.86	2:17	1.99	1:22:26	71.93	5:59	5.22
25	15:29	12.17	2:25	1.90	1:42:04	80.24	7:14	5.69
26	27:20	22.89	2:13	1.85	1:24:05	70.42	6:15	5.23
Odd	17:33	15.10	2:23	2.13	1:30:55	78.12	5:23	4.62
Even	23:42	20.52	2:14	1.92	1:26:15	73.40	4:49	4.15
Total	20:38	17.81	2:19	2.02	1:28:35	75.76	5:06	4.38

Time* = Normalized time = $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

o FORCON Includes CACOPY from this point on

oo Output Includes PRLXT from this point on

Table 10-5
PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

Issue	Format Conversion Time*	%	Input Time*	%	Search Time*	%	Output Time*	%
1	17:41	16.11	2:26	2.21	1:19:05	72.03	10:36	9.65
2	29:04	24.53	2:14	1.88	1:20:24	67.85	6:48	5.74
3	14:02	11.84	2:23	2.01	1:35:22	80.48	6:43	5.67
4	27:21	21.81	2:18	1.83	1:29:31	71.38	6:15	4.98
5	19:45	15.56	2:00	1.57	1:38:22	77.52	6:11	4.87
6	24:23	23.49	2:20	2.25	1:11:49	69.18	5:16	5.08
7	14:27	13.24	2:29	2.27	1:25:54	78.67	6:21	5.82
8	34:51	22.78	2:44	1.79	1:47:33	70.29	7:52	5.14
9	22:57	16.93	2:30	1.84	1:43:19	76.19	6:49	5.03
10	34:08	23.51	1:40	1.15	1:41:44	70.06	7:40	5.28
11	22:57	16.96	2:36	1.92	1:42:09	75.50	7:36	5.62
12	31:59	21.58	2;26	1.64	1:45:52	71.43	7:56	5.35
13	19:12	11.23	2:43	1.59	2:22:06	83.10	6:59	4.08
14	43:15	23.10	2:41	1.43	2:11:10	70.70	10:07	5.40
15	25:56	10.86	2:42	1.13	3:21:16	84.28	8:54	3.73
16	34:40	19.36	2:42	1.51	2:13:49	74.72	7:54	4.41
17	27:34	14.05	2:39	1.35	2:37:39	80.35	8:20	4.25
18	28:06	17.67	2:42	1.70	2:00:53	76.03	7:19	4.60
19	29:36	14.34	2:43	1.32	2:43:50	79.38	10:14	4.96
20	28:16	16.59	2:46	1.62	2:11:38	77.25	7:43	4.53
21	26:32	14.13	2:45	1.46	2:29:09	79.42	9:22	4.99
22	33:54	17.52	2:48	1.45	2:27:22	76.16	9:25	4.87
23	28:00	15.25	2:40	1.45	2:23:32	78.18	9:24	5.12
24	43:54	19.03	2:53	1.25	2:53:51	75.36	10:04	4.36
25	38:39	17.13	2:40	1.18	2:52:24	76.42	11:52	5.26
26	45:26	20.63	3:02	1.38	2:40:59	73.11	10:43	4.87
Odd	23:38	14.43	2:34	1.64	2:10:19	78.58	8:05	5.31
Even	33:48	20.89	2:35	1.61	1:59:44	72.53	8:05	4.97
Total	28:43	17.66	2:35	1.63	2:04:58	75.56	8:05	5.14

$$\text{Time* = Normalized time} = \frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$$

Table 10-6

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

Issue	Format Conversion Time*	%	Input Time*	%	Search Time*	%	Output Time*	%
1	12:53	7.40	2:45	1.58	2:27:47	84.93	10:36	6.09
2	15:40	9.76	2:43	1.69	2:11:44	82.08	10:25	6.49
3	11:36	6.59	2:03	1.16	2:31:10	85.84	11:16	6.40
4	17:58	10.24	2:02	1.16	2:22:01	80.92	13:29	7.68
5	14:10	8.06	2:30	1.42	2:27:20	83.81	11:48	6.71
6	15:59	9.92	2:45	1.71	2:10:40	81.11	11:42	7.26
7	10:42	8.02	2:35	1.93	1:51:13	83.31	9:00	6.74
8	17:25	9.84	2:44	1.54	2:25:25	82.16	11:27	6.47
9	12:02	7.44	2:37	1.62	2:16:31	84.43	10:31	6.50
10	14:05	9.95	2:45	1.94	1:54:43	81.02	10:02	7.08
11	11:46	7.25	2:32	1.56	2:16:55	84.36	11:04	6.82
12	13:17	10.25	2:29	1.91	1:44:20	80.50	9:31	7.34
13	9:40	9.68	2:22	2.37	1:19:04	79.14	7:45	7.76
14	9:36	11.77	2:20	2.85	1:03:05	77.31	6:34	8.05
15	9:42	9.31	3:23	3.25	1:22:46	79.51	8:15	7.93
16	12:54	11.62	2:31	2.26	1:28:02	79.31	7.34	6.81
17	12:10	10.96	2:29	2:23	1:27:35	78.90	8.47	7.91
18	10:01	10.84	2:33	2.76	1:13:01	79.02	6.43	7.27
19	6:30	10:82	2:17	3.81	0:44:00	73.34	7:13	12.03
20	15:37	11.86	2:36	1.97	1:42:44	78.00	10:46	8.17
21	8:00	10.84	2:11	2.96	0:57:01	77.26	6:36	8.94
22	13:20	11.89	2:17	2.03	1:27:05	77.61	9:30	8.47
23	8:59	10.85	2:24	2.89	1:04:12	77.54	7:12	8.70
24	13:32	11.39	2:28	2.07	1:32:22	77.75	10:27	8.79
25	9:45	11:13	2:17	2.61	1:07:01	76.50	8.33	9.76
26	12:57	11.63	1:56	1.74	1:24:54	76.28	11:31	10.34
Odd	10:37	9:10	2:30	2.26	1:40:58	80.68	9:07	7.87
Even	14:01	10:84	2:28	1.97	1:30:47	79.47	9:59	7.71
Total	12:19	9.97	2:29	2.12	1:35:53	80.75	9:32	7.79

$$\text{Time* = Normalized time} = \frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$$

Table 10- 7

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	7:22	11.12	1:59	2.99	50:18	75.92	6:36	9.97
2	10:15	13.88	2:01	2.73	54:21	73.64	7:11	9.73
3	7:12	8.95	2:17	2.13	1:06:57	83.25	3:59	4.95
4	9:35	9.00	2:21	2.21	1:24:20	79.17	10:15	9.61
5	7:27	9.98	2:23	2.95	59:48	79.72	5:21	7.35
6	7:54	8.99	3:16	3.71	1:08:23	77.81	8.20	9.48
7	6:54	9.52	2:22	3.26	57:13	78.92	6:02	8.02
8	12:49	10.29	2:31	2.02	1:39:43	80.17	9:20	7.51
9	8:02	9.66	2:22	2.84	1:05:24	78.65	7:21	8.84
10	13:15	9.88	2:17	1.70	1:44:22	77.87	14:08	10.54
11	8:46	8.20	2:43	2.54	1:26:06	80.53	9:20	8.72
12	13:02	10.22	2:14	1.75	1:41:12	79.39	11:00	8.63
13	8:18	9.84	2:20	2.76	1:05:54	78.14	7:48	9.25
14	12:34	9.57	2:19	1.76	1:43:46	79.10	12:32	9.55
15	8:17	8.54	3:58	4.09	1:16:32	78.90	8:13	8.47
16	12:32	9.04	3:27	2.52	1:47:58	78.76	13:17	9.68
17	8:33	8.62	3:12	3.23	1:18:20	79.00	9:05	9.16
18	12:00	11.98	2:33	2.54	1:13:07	73.00	12:30	12.48
19	9:03	12.20	2:36	3.50	52:30	70.79	10:01	13.52
20	12:04	12.96	2:35	2.56	1:10:38	70.01	15:36	15.43
21	9:13	12.31	2:30	3.34	53:12	71.09	9:55	13.25
22	12.33	12.80	2:14	2.28	1:10:41	72.06	12:37	12.86
23	9:01	12.71	2:05	2.93	49.55	70.37	9:55	13.98
24	12:29	13.16	2:28	2.59	1:07:23	70.52	13:13	13.83
25	10:22	16.08	2:12	3.41	42:24	65.75	9:31	14.76
26	13:54	13.41	2:28	2.39	1:11:42	69.18	15:35	15.03
Odd	8:17	10.16	2:35	3.17	1:02:56	77.16	7:46	9.52
Even	11:40	10.78	2:23	2.34	1:25:19	76.45	11:31	10.50
Total	9:59	10.37	2:29	2.58	1:14:08	77:02	9:39	10.03

$$\text{Time* = Normalized time} = \frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$$

Table 10-8

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

BIORESEARCH INDEX VOLUME 71

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	8:15	6.95	1:04	0.91	1:47:07	90.17	3.09	2.65
2	7:59	6.25	0:00	0.00	1:57:31	91.95	2:18	1.80
3	7:47	7.77	0:55	0.92	1:29:29	89.31	2:00	2.00
4	8:05	8.52	1:01	1.07	1:22:34	87.09	3:09	3.32
5	8:13	7.58	1:02	0.96	1:35:39	88.32	3:31	3.24
6	8:38	7.53	0:00	0.00	1:42:13	89.19	3:46	3.28
7	7:48	4.53	0:00	0.00	2:37:10	91.27	7:13	4.19
8	8:49	7.83	1:10	1.04	1:36:54	86.13	5:38	5.00
9	8:12	6.83	1:07	.93	1:44:38	87.20	6:04	5.05
10	8:10	7.93	1:15	1.21	1:27:48	85.32	5:42	5.54
11	7:51	6.27	1:04	0.85	1:51:26	89.07	4:47	3.82
12	8:15	6.86	0:00	0.00	1:45:51	87.99	6:11	5.14
Total	8:10	7.07	0:43	0.66	1:44:52	88.58	4:27	3.75

$$\text{Time* = Normalized Time} = \frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$$

Table 10-9

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

BIOLOGICAL ABSTRACTS VOLUME 51

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1								
2								
3								
4			(Data for 1-6 do not exist)					
5								
6								
7	6:00	14.30	0:52	2.08	33:58	80.89	1:09	2.73
8	6:04	15.54	0:00	0:00	31:52	81.69	2:11	3.61
9	5:18	17.66	0:00	0:00	23:43	79.06	0:59	3.28
10	5:19	7.52	0:57	1.35	1:03:10	89.22	1:21	1.91
11	5:19	7.91	1:05	1.61	59:27	88.47	1:15	1.88
12	5:26	8.32	0:47	1.21	57:40	88.17	1:30	2.30
13	5:57	9.98	0:56	1.57	51:30	86.26	1:18	2.19
14	5:25	9.39	1:50	3.19	49:58	86.76	1:15	2.18
15	5:46	8.66	1:01	1.69	52:03	87.18	1:28	2.47
16	5:21	10.38	0:59	1.89	43:32	84.37	1:44	3.36
17	5:03	9.20	0:58	1.77	47:01	85.65	1:51	3.38
18	5:15	9.77	1:00	1.87	45:44	85.17	1:42	3.19
19	5:14	6.38	1:04	1.30	1:13:43	90.01	1:54	2.31
20	5:01	6.69	0:49	1.08	1:07:23	89.85	1:47	2.38
21	5:02	6.69	1:06	1.47	1:05:11	86.57	3.58	5.27
22	4:58	6.23	1:08	1.42	1:10:33	88.19	3.24	4.26
23	5:09	5.96	1:12	1.40	1:16:46	88.85	3.16	3.79
24	4:52	6.32	0:00	0.00	1:09:09	89.68	3:05	4.00
Total	5:22	9.27	0:52	1.38	54:34	86.45	1:57	3.03

Time* = Normalized Time = $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

Table 10-10

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

BIOLOGICAL ABSTRACTS VOLUME 52

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	5:00	5.85	0:00	0.00	1:16:59	89.72	3:48	4.43
2	5:14	5.14	1:10	1.15	1:31:48	90.26	3:29	3.43
3	5:14	6.06	1:04	1.23	1:17:02	89.15	3:05	3.56
4	5:02	5.25	1:06	1.15	1:26:46	90.39	3:06	3.23
5	5:54	7.48	0:54	1.14	1:10:04	88.80	2:02	2.58
6	5:55	7.47	0:40	0.84	1:10:27	88.96	2:10	2.73
7	6:10	7.47	0:56	1.14	1:13:25	88.99	1:59	2.40
8	5:56	7.28	0:00	0.00	1:13:28	90.24	2:01	2.48
9	6:13	7.37	0:59	1.16	1:14:32	88.42	2:34	3.05
10	5:59	6.40	0:00	0.00	1:24:44	90.81	2:36	2.79
11	6:42	6.12	1:02	0.95	1:36:48	88.40	4:57	4.52
12	6:31	5.39	0:00	0.00	1:48:15	89.53	6:39	5.09
13	6:03	5.93	1:18	1.27	1:28:37	86.88	6:02	5.92
14	6:19	5.43	0:00	0.00	1:43:58	89.32	6:06	5.24
15								
16	6:34	5.89	0:50	.75	1:38:35	88.34	5:19	4.76
17	6:21	6.37	1:06	1.11	1:26:59	87.33	5:10	5.18
18	6:36	6.06	0:00	0.00	1:37:12	89.26	5:05	4.67
19	5:23	7.50	0:00	0.00	1:01:45	86.12	4:34	6.37
20	6:40	6.19	0:00	0.00	1:35:47	88.93	5:15	4.88
21	6:39	5.89	1:13	1.08	1:39:28	88.18	5:29	4.86
22	6:09	5.63	1:09	1.05	1:36:19	88.20	5:29	5.02
23	6:06	6.04	0:00	0.00	1:29:54	88.93	5:05	5.02
24	6:21	6.20	1:04	1.04	1:30:31	88.48	4:23	4.28
Total	6:03	6.28	0:38	0.65	1:26:41	88.84	4:11	4.20

Time* = Normalized Time = $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

Table 10-11

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS. ISSUE

ENGINEERING INDEX VOLUME 71

	Format Conversion		Input		Search		Output	
Issue	Time*	%	Time*	%	Time*	%	Time*	%
1	7:01	43.35	0:27	2.79	7:14	44.68	1:29	9.18
2	17:44	47.30	6:14	16.60	4:48	12.80	8:44	23.3
3	9:12	40.34	0:33	2.42	10:17	45.11	2:46	12.12
4°								
5°								
6	9:28	38.00	0:00	0.00	12:22	49.67	3:04	12.33
7	9:31	34.50	0:34	2.05	13:40	49.54	3:50	13.90
8	9:33	33.50	0:36	2.13	14:25	50.59	3:56	13.78
9	9:45	31.53	0:00	0.00	16:45	54.19	4:25	14.27
10	9:22	24.97	0:44	1.96	20:36	54.94	6:48	18.14
11	9:34	31.25	0:38	2.06	15:58	52.20	4:26	14.49
12	9:03	11.47	1:31	1.93	59:08	74.95	9:12	11.65
Total		33.62		3.19		48.87		14.32

Time* = Normalized Time = $\frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$

° Data for issues 4-5 do not exist

Table 10-12

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

ENGINEERING INDEX VOLUME 72

Issue	Format Conversion Time*	%	Input Time*	%	Search Time*	%	Output Time*	%
1	7:53	17.39	1:16	2.81	30:31	67.37	5:38	12.43
2	4:41	10.46	1:02	2.32	34:16	76.67	4:43	10.54
3	5:41	12.88	1:33	3.53	31:45	72.00	5:07	11.59
4	5:41	10.83	1:18	2.47	39:17	74.65	6:20	12.05
5	7:47	12.36	1:01	1.62	46:5[illegible]	74.57	7:13	11.45
Total	6:10	12.78	1:14	2.55	36:32	73.05	5:48	11.61

$$\text{Time* = Normalized Time} = \frac{\text{Actual Cost of Operation}}{\text{cpu charge}}$$

Table 10-13

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM
VS ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

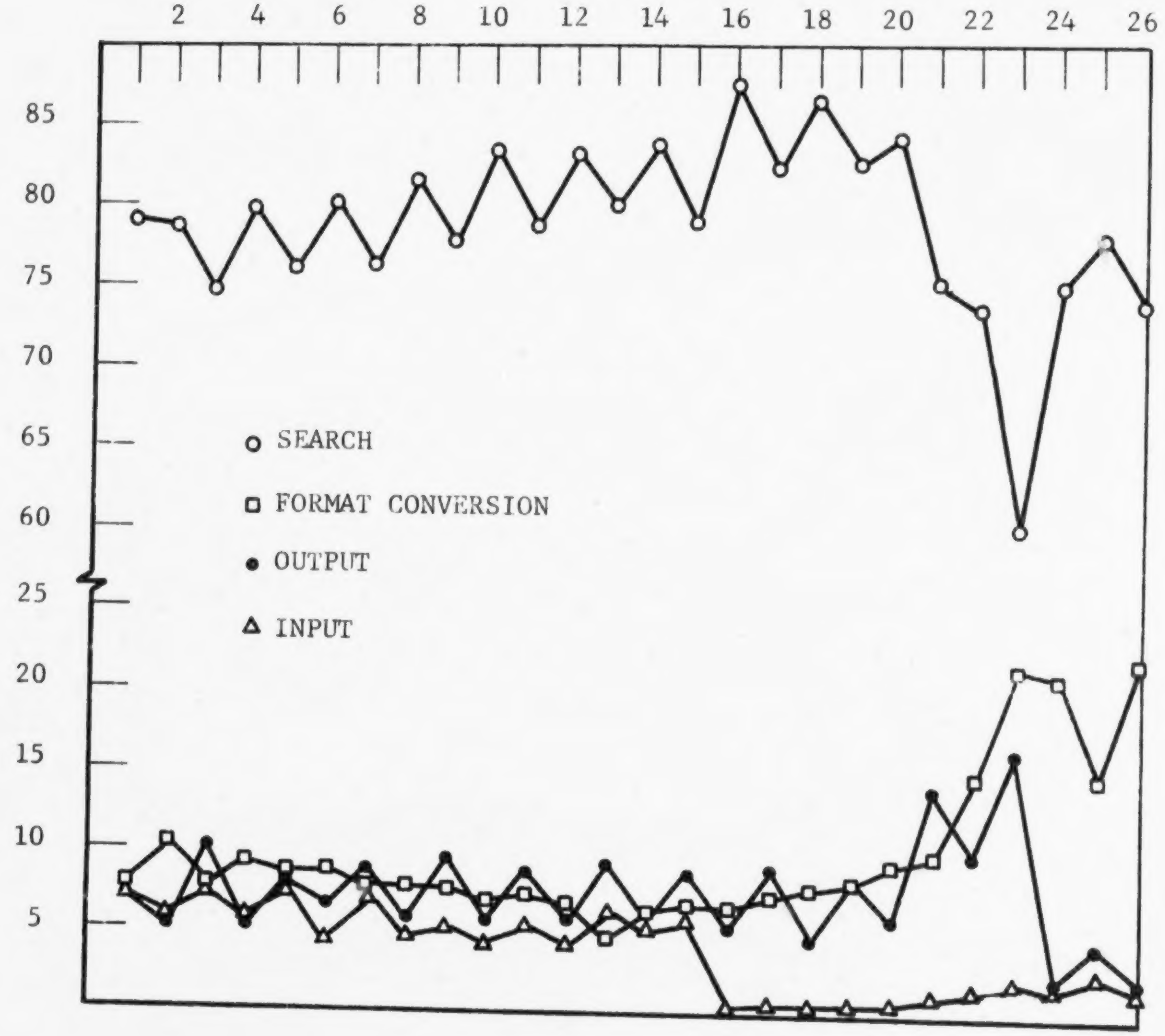


Figure 10-16
PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

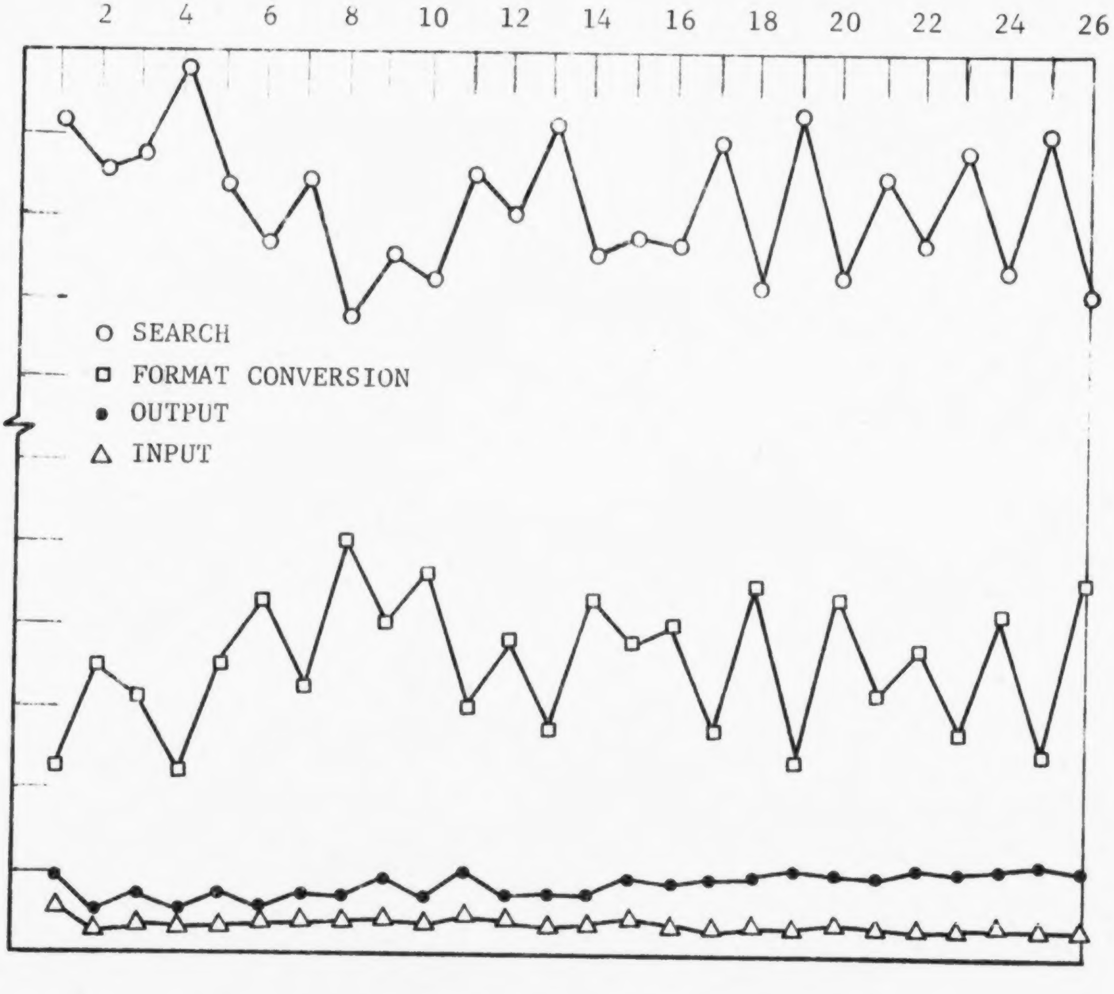


Figure 10-17
PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

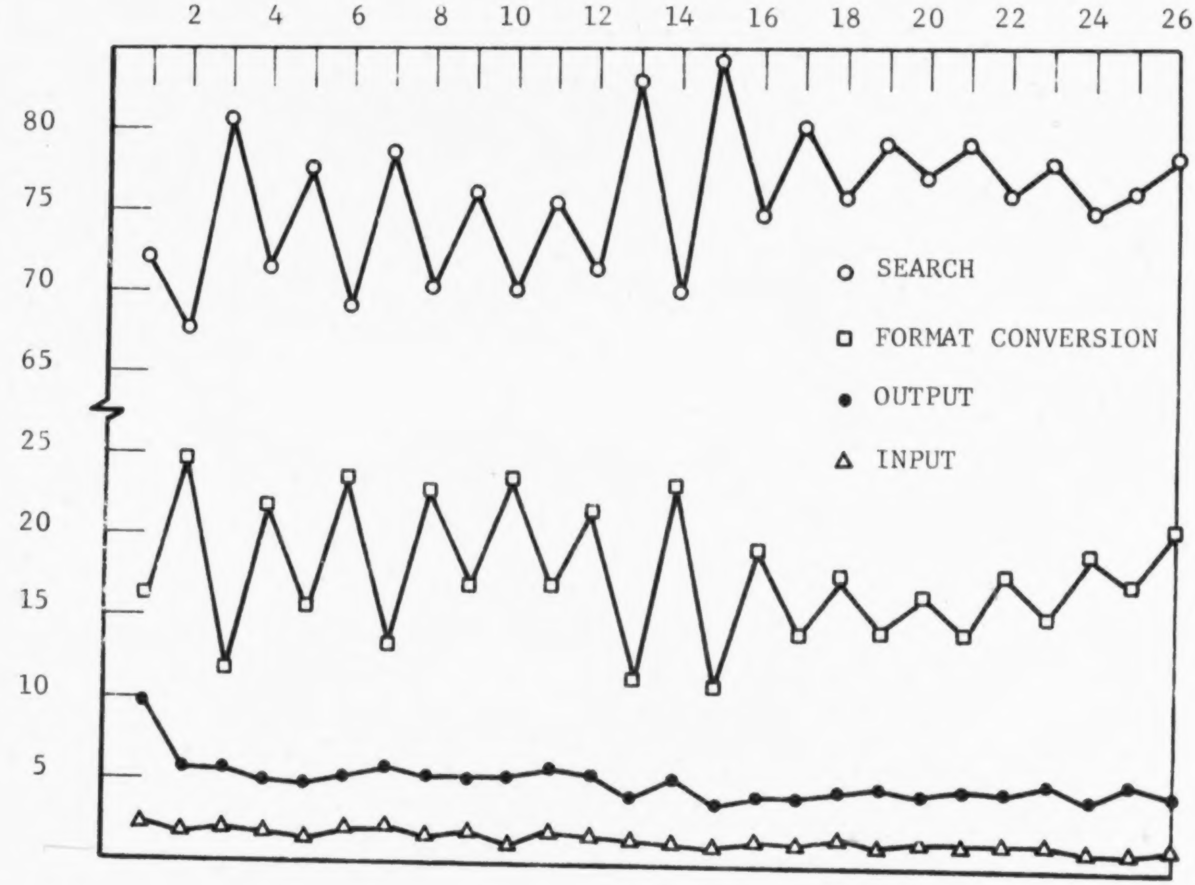


Figure 10-18

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

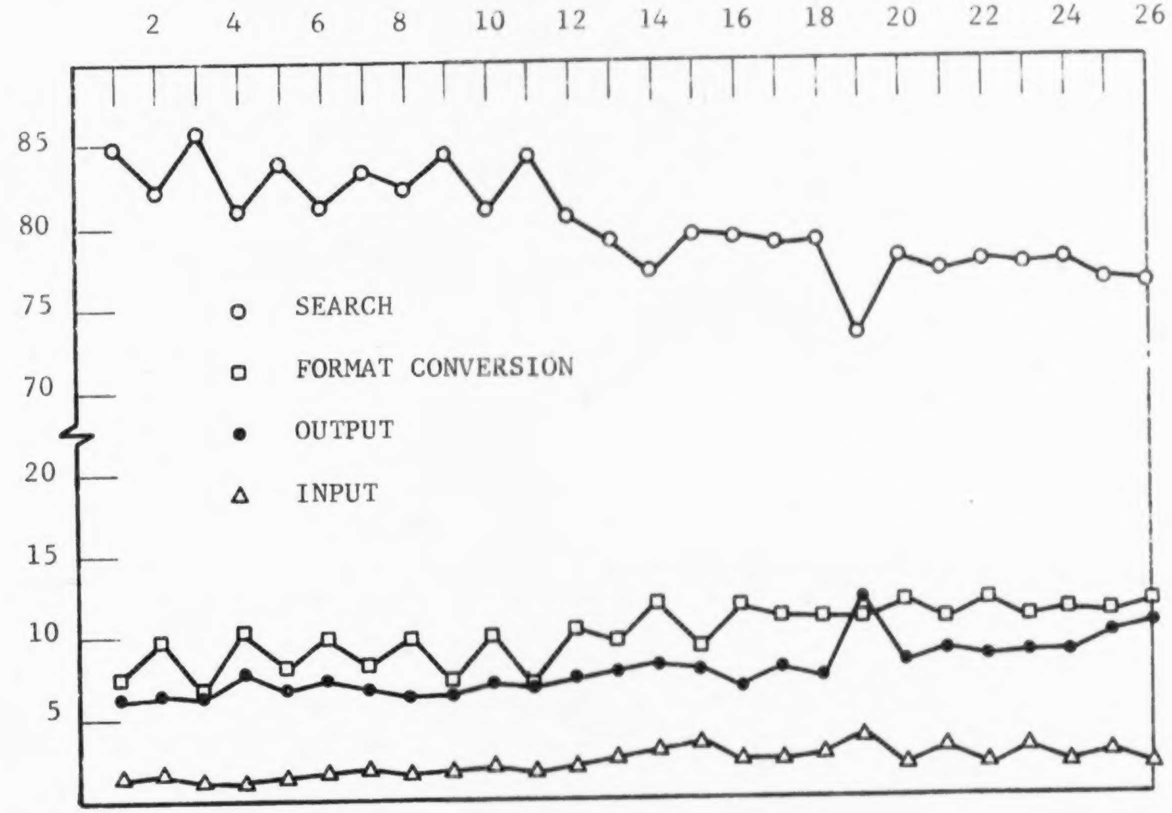


Figure 10-19

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

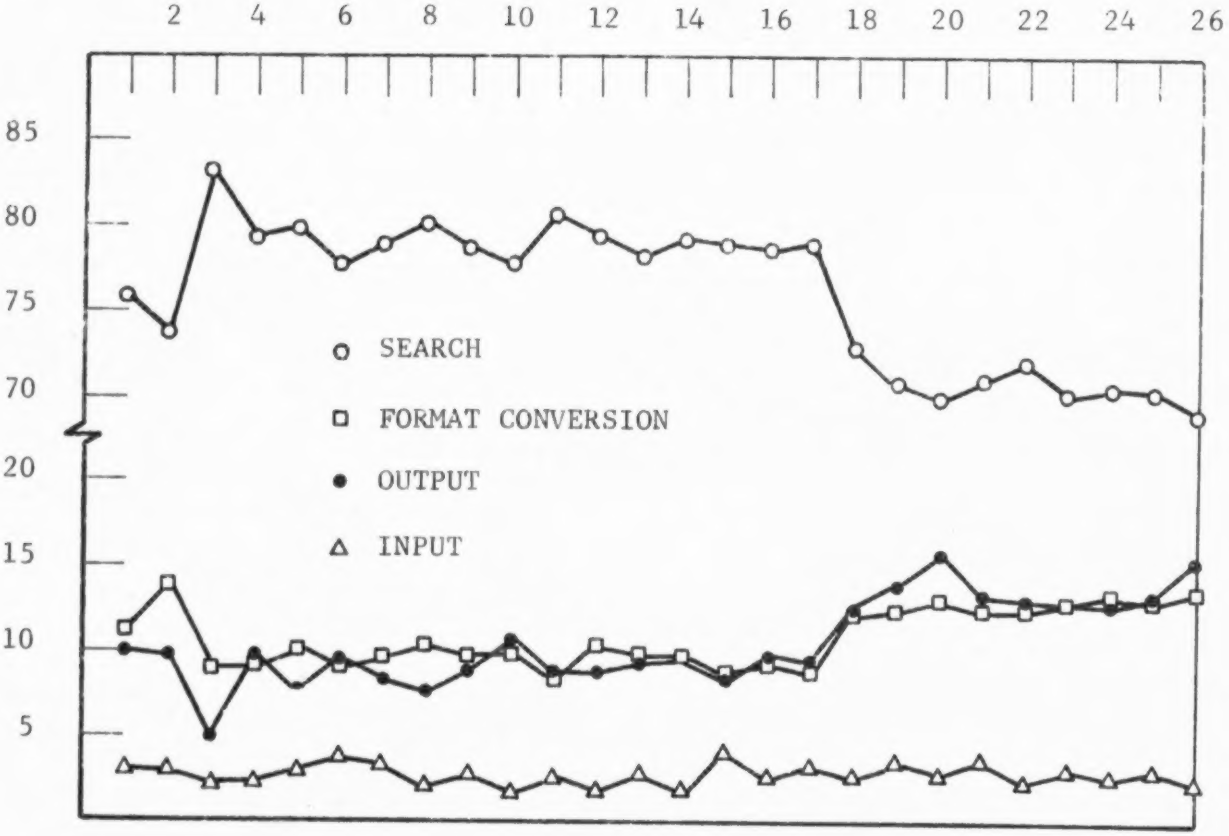


Figure 10-20

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

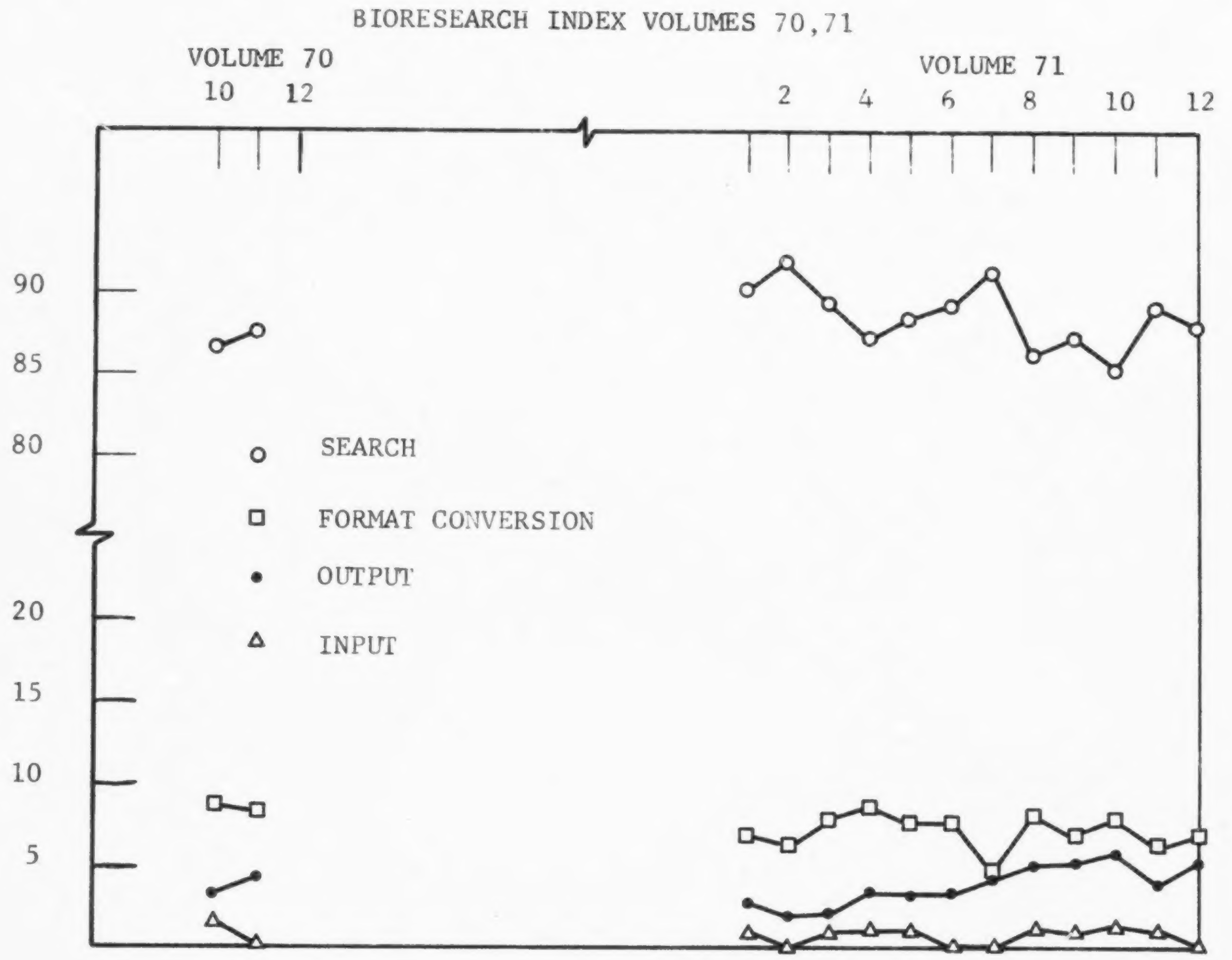


Figure 10-21

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

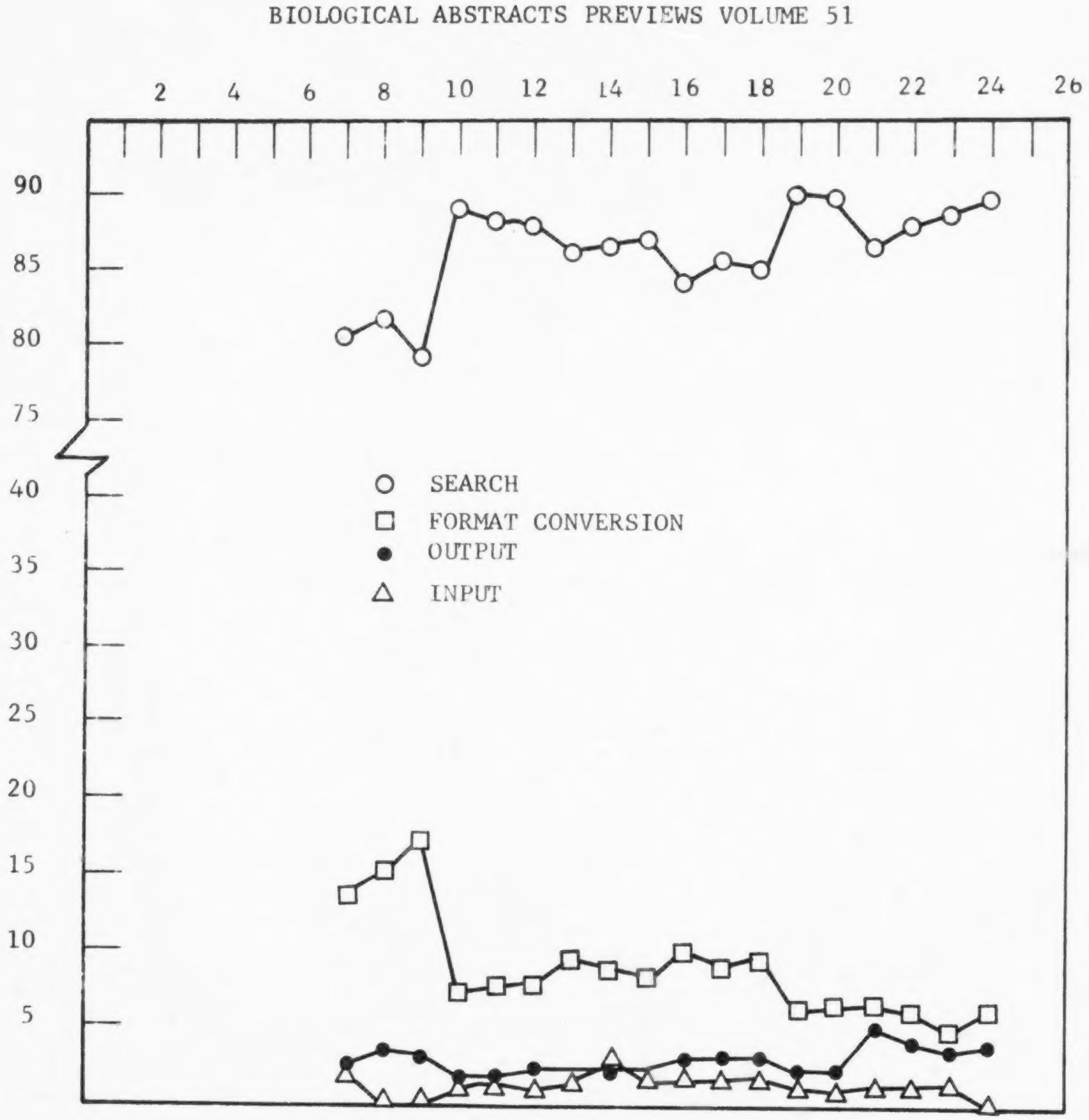


Figure 10-22

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

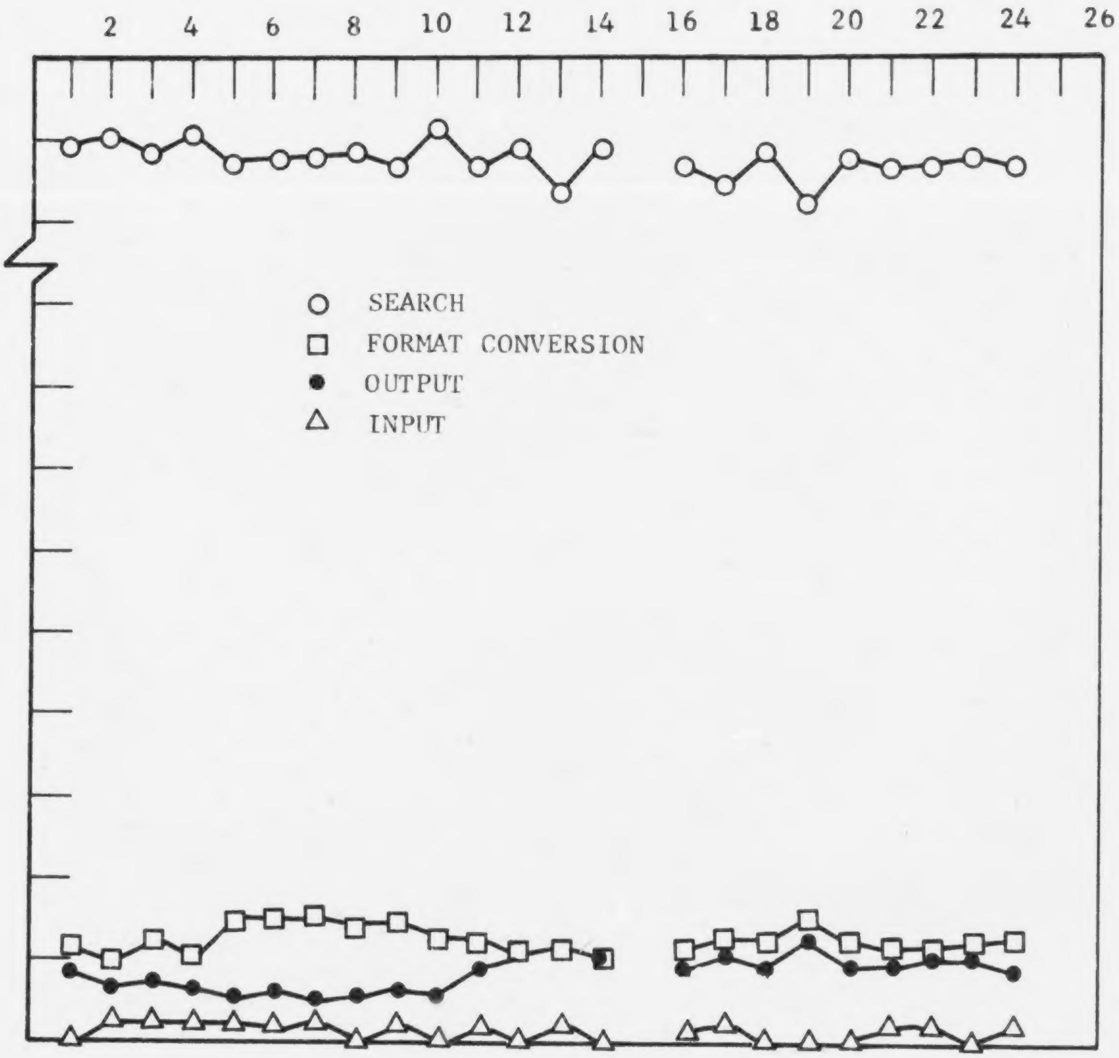


Figure 10-23

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71,72

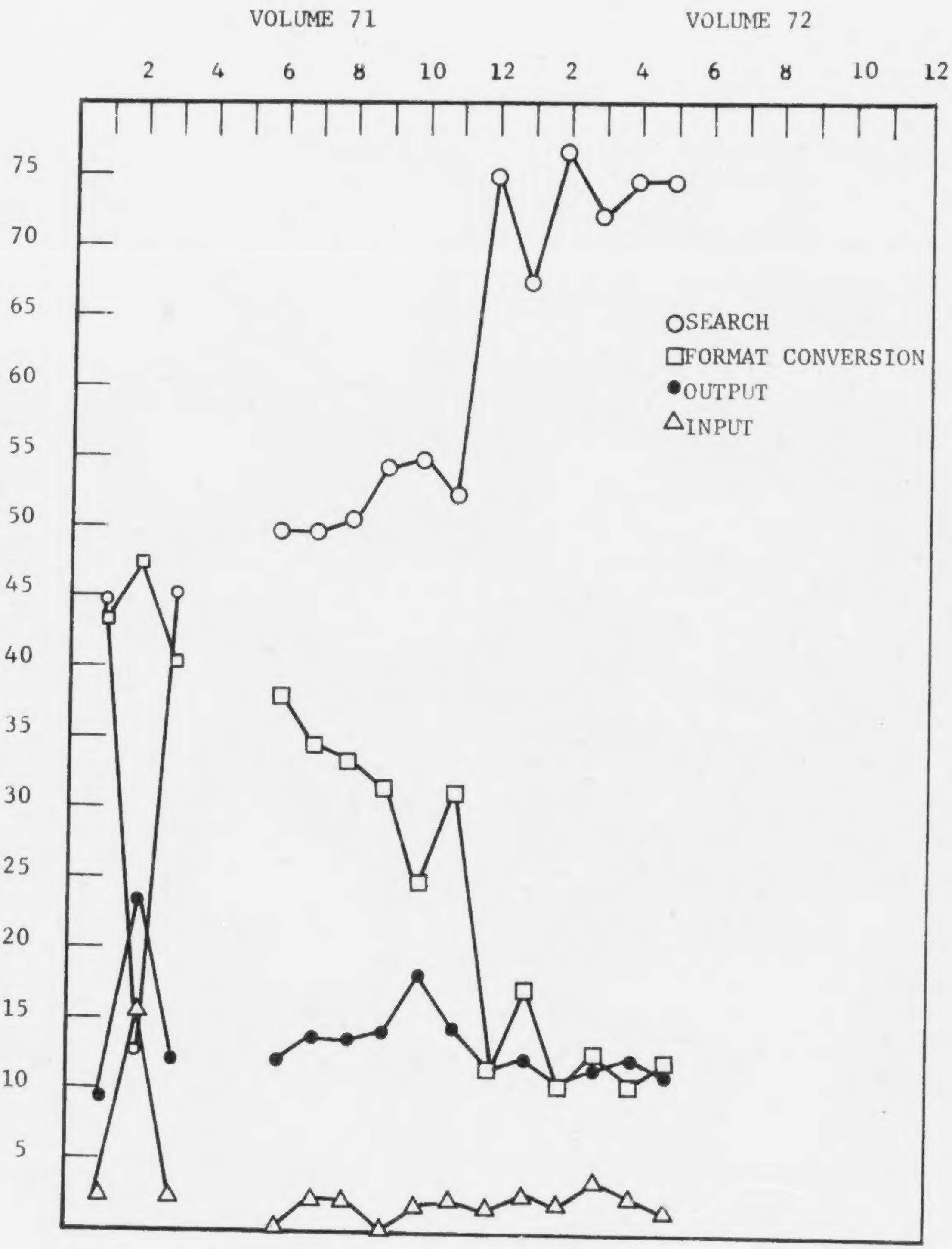


Figure 10-24

PERCENT OF COMPUTER TIME PER COMPUTER PROGRAM VS. ISSUE

The cost of running the programs is naturally dependent on the size and number of input profiles, the data base, the number of terms, the frequency with which the terms appear in the issue searched, the number of citations in the issue searched, the number of near hits (i.e., citations that matched profile terms but were subsequently disqualified on the basis of logic or weights etc.), the number of hits obtained by the profiles, and the number of citations printed.

The two most significant determinants of cost are the size of the data base and the number of profile terms. CSC has developed a formula whereby given the number of profile terms and the number of citations in a data base we can predict the cost of the run to within 10%. The cost for searching one profile term against one CA Condensates citation is $\$1.05 \times 10^{-5}$ based on a total term list of 500-5000 words. The CSC Constant Computer Cost Factor is:

$$\$1.05 \times 10^{-5}/\text{profile-term/citation.}$$

The search portion of the system is the prime determinant--other factors such as number of hits, complexity of logic, number of hits printed,etc. account for the 10% variation.

10.4.2 CSC Time and Cost Summary Sheet

The CSC Time and Cost Summary Sheet is prepared for each issue of each data base processed. It is color coded for CA (white), BA (green), and EI (yellow). These are attached as Figures 10-25, 10-26, and 10-27. The time figures recorded on these sheets are obtained from the computer printout listings for each of the programs. A sample of the printout listing of the INPUTR program for a production run of CA Condensates is attached as Figure 10-28. Percentage of total time and cost figures are calculated. The cost figure is obtained from CPU time, core size and current computer rates. The statistics calculated on the following page require input from the computer-generated Production Run Summary--Computer Search Center, which is discussed below. The Time and Cost Summary contains the following:

- date, data base, volume and issue
- time in seconds and in hours, minutes and seconds for each program (and any reruns that are necessary)
- percentage of total time used by each program
- cost per program
- total time and cost for all programs and reruns
- time and cost per profile
- time and cost per term (profile term)
- time and cost per hit
- time and cost per term/per citation
- cost average to data (begun anew with each volume)

Statistics Recorded

Date of Run	________
Tape Service	CA CONDENSATES
Volume - Issue	__:__

PROGRAM	SEC.	HH:MM:SS	%	COST	RERUNS NO.	RERUNS TOTAL TIME
CACOPY		: :				
FORCON		: :				
DXEDIT		: :				
INPUTR		: :				
SEARCH		: :				
CACARD		: :				
STIXA		: :				
OCP		: :				
PRINT		: :				
PRILIB		: :				
TOTAL		: :	100.00	$		

Additional Cost due to Reruns $______ $______ Total Cost

Statistics Calculated

Time & Cost per Profile	______ sec.	$ ______
Time & Cost per Term	______ sec.	$ ______
Time & Cost per Hit	______ sec.	$ ______
Time & Cost per Term per Citation	______ sec.	$ ______
Cost Average to Date		$ ______

Figure 10-25

COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET

Statistics Recorded

Date of Run ____________
Tape Service BA PREVIEWS
Volume - Issue __:__

PROGRAM	SEC.	HH:MM:SS	%	COST	RERUNS	
					NO.	TOTAL TIME
BACOPY		: :				
FORBAP		: :				
DKEDIT		: :				
INPUTR		: :				
BASRCH		: :				
BAPFORM		: :				
STIXA		: :				
OCP		: :				
PRINT		: :				
PRILIB		: :				
TOTAL		: :	100.00	$		

Additional Cost due to Reruns $__________ $__________ Total Cost

Statistics Calculated

Time & Cost per Profile __________ sec. $__________
Time & Cost per Term __________ sec. $__________
Time & Cost per Hit __________ sec. $__________
Time & Cost per Term per Citation __________ sec. $__________

Cost Average to Date $__________

Figure 10-26
COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET

Statistics Recorded

Date of Run ____
Tape Service EI COMPENDEX
Volume - Issue __:__

PROGRAM	SEC.	HH:MM:SS	%	COST	RERUNS NO.	RERUNS TOTAL TIME
EICOPY		: :				
EICON		: :				
DKEDIT		: :				
INPUTR		: :				
EISRCH		: :				
EICARD		: :				
STIXA		: :				
EIOCP		: :				
PRINT		: :				
PRILIB		: :				
TOTAL		: :	100.00	$		

Additional Cost due to Reruns $____ $____ Total Cost

Statistics Calculated

Time & Cost per Profile ____ sec. $____
Time & Cost per Term ____ sec. $____
Time & Cost per Hit ____ sec. $____
Time & Cost per Term per Citation ____ sec. $____

Cost Average to Date $____

Figure 10-27
COMPUTER SEARCH CENTER TIME AND COST SUMMARY SHEET

```
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR
STAT-TAB  HASP 3.0   START JOB 920    9.02.41 AM 23 JUN 72   ACCT 80351   B2810500   CA-PRODN INPUTR

                                          H A S P   S Y S T E M   L O G

$ 23.28.59 JOB 920 -- B2810500 -- BEGINNING EXEC - INIT 1 - CLASS F
 *23.29.58 JOB 920            80351 B2810500.LKED      STEP TIME=     3.27 SEC CORE=299,USED=093
 *23.30.08 JOB 920 IEF233A M 281,IS1660,,B2810500,GO,CALCBS
 *23.30.08 JOB 920 IEF233A M 282,IS1632,,B2810500,GO,PCSCCARD
 *23.30.57 JOB 920 IEC114E D 281
 *23.30.58 JOB 920 IEC101A M 281,IS1660,SL,B2810500,GO,CALCBS
 *23.38.49 JOB 920 IEC209I B2810500 IS1660 281 TR=000,TW=000,EG=000,CL=000,N=000,SIO=00041
 *23.38.49 JOB 920 IEC209I B2810500 IS1632 282 TR=000,TW=000,EG=000,CL=000,N=000,SIO=00058
 *23.39.03 JOB 920 IEF280E K 281,IS1660,B2810500,GO,CALCBS
 *23.39.03 JOB 920 IEF280E K 282,IS1632,B2810500,GO,PCSCCARD
 *23.39.03 JOB 920            80351 B2810500.GO        STEP TIME=    98.59 SEC CORE=299,USED=299
 *23.39.04 JOB 920            80351 B2810500.          JOB
$ 23.39.07 JOB 920 HASP04CA ACCTG/DA-ACCESS DISCREPANCY
$ 23.39.07 JOB 920 DONE           3,636 LINES
$ 23.39.07 JOB 920 DONE
```

101.86

Figure 10-28

TIME RECORD FROM CA PRODUCTION INPUTR RUN

10.4.3 Production Run Summary--CSC

The Production Run Summary--CSC is a machine generated summary of statistics for each issue of each data base. It includes three sections. The first page (see Figure 10-29) contains:

- date, data base, volume, issue and number of citations
- number of profiles and number of in-house profiles
- number and mean of input terms
- number and mean of aggregated terms
- percent term reduction by aggregation
- number of hits (in-house, others and total) and means, both recorded and printed
- number and mean of unique citations retrieved
- number and mean of cards printed
- hits recorded and hits printed per citation retrieved
- range and median of hits generated
- range of hits printed
- number of profiles getting no hits

The second page (more than one page is printed if necessary) gives the distribution of hits by profile (see Figure 10-30). The third page (again more than one page is printed if necessary) gives the distribution of hits recorded and printed by corporate code (see Figure 10-31).

10.4.4 Profile Term Statistics

Another listing prepared for each issue of each data base is shown in Figure 10-32. It gives statistics on term processing for all the terms in all the profiles. It includes:

- number of input terms
- number of unique terms
- number of Least Common Bigrams (LCB's) found in the terms

```
                PRODUCTION RUN SUMMARY-----COMPUTER SEARCH CENTER

                   DATE OF RUN          JUNE 17, 1972

        SERVICE, VOLUME, ISSUE          CA CONDENSATES 76:25

             CITATIONS ON TAPE          5541
------------------------------          --------------------------
              PROFILES IN RUN            129          ( 31 IN-HOUSE)

        NUMBER OF INPUT TERMS           3458          (26.8/PROFILE)

   NUMBER OF AGGREGATED TERMS           2595          (20.1/PROFILE)

PERCENT REDUCTION BY AGGREGATION         25.00
------------------------------          -----------  -------------
      HITS RECORDED

                 .....IN-HOUSE           677          (21.8/PROFILE)

                 .......OTHERS          3962          (40.4/PROFILE)

                 ........TOTAL          4639          (35.9/PROFILE)

      HITS PRINTED

                 .....IN-HOUSE           677          (21.8/PROFILE)

                 .......OTHERS          3711          (37.8/PROFILE)

                 ........TOTAL          4388          (34.0/PROFILE)
------------------------------          --------------------------
   UNIQUE CITATIONS RETRIEVED           2819          (21.8/PROFILE)

                 CARDS PRINTED          4646          (36.0/PROFILE)

HITS RECORDED/CITATION RETRIEVED           1.64

HITS PRINTED/CITATION RETRIEVED            1.55
------------------------------          --------------------------
                RANGE OF HITS            0 - 383

                       MEDIAN             19.0

              RANGE OF PRINTS            0 - 369

    NUMBER OF ZERO-HIT PROFILES           11
```

Figure 10-29

PRODUCTION RUN SUMMARY--OVERALL SUMMARY

```
          PRODUCTION RUN SUMMARY-----COMPUTER SEARCH CENTER

JUNE 17, 1972                                     CA CONDENSATES 76:25

                       PROFILE HIT DISTRIBUTION

 NUMBER       NUMBER                  NUMBER       NUMBER
OF  HITS    OF PROFILES              OF  HITS    OF PROFILES

    0           11                      27            4
    1            6                      29            2
    2            7                      30            2
    3            5                      31            2
    4            7                      32            1
    5            3                      35            1
    6            2                      37            1
    7            4                      39            2
    8            1                      41            2
    9            2                      42            2
   11            2                      44            1
   13            4                      46            1
   14            1                      47            2
   15            2                      49            2
   16            4                      52            1
   17            2                      54            1
   18            1                      58            1
   19            2                      59            1
   20            1                      60            1
   21            1                      73            1
   22            2                      76            1
   23            3                      77            2
   24            2                      86            1
   25            2                      88            3
   26            1                      93            1
```

Figure 10-30
PRODUCTION RUN SUMMARY--PROFILE HIT DISTRIBUTION

```
        PRODUCTION RUN SUMMARY-----COMPUTER SEARCH CENTER

JUNE 17, 1972                              CA CONDENSATES 76:25

              CORPORATE DISTRIBUTION OF HITS AND PRINTS

 CODE      HITS     PRINTS        CODE      HITS     PRINTS

  A01       169       169          L24       136       136
  A05        24        24          L25         4         4
  A07       120       120          L26        22        22
  A10       147       147          L28        78        78
  A13       140       140          L29        24        24
  A16         4         4          L35         5         5
  A23       502       319          L39         0         0
  A24        31        31          L41         7         7
  A25         9         9          L44        30        30
  A26         0         0          L45        19        19
  A27       164       164          L48        23        23
  G01        70        70          L49        79        79
  G06        71        71          L55        13        13
  G07       123       123          L56        75        75
  G11        31        31          L58        24        24
  L01       446       446          L59       100       100
  L02         2         2          L60         1         1
  L05        52        50          L62        47        47
  L06        69        69          L63        35        35
  L09        90        90          L64        10        10
  L10       539       539          W04       220       220
  L11        56        56          W06       117       117
```

Figure 10-31

PRODUCTION RUN SUMMARY--CORPORATE HIT DISTRIBUTION

```
2699    2     022 COMPONENT                         0    2
2700    2     022 PHASE                             0    0
2701    2-    C22-COMPONENT                         0    2
2702    2-    C22-PHASE                             0    0
2703    4-    021,2,4-TRIAZOL                       4    3
2704    4-    022,4-D                               2    0
2705    49    C2049CCO                              1    1
2706    5-    C22,4,5-T                             4    0
2707    76    C2076CCO                              1    1
2708    79    02079CCO                              1    1


STEP TWO COMPLETE.


RESULTS OF TERM PROCESSING

 3747 TERMS
 2709 UNIQUE TERMS
  465 LCBS USED (OUT OF  2003 )

MEAN FREQ. OF TERM LCBS IS     19453.009
S.D. OF FREQ. OF TERM LCBS IS 11079.864

MEAN FREQ. OF ALL LCBS IS       8410.566

MEAN GROUP SIZE IS       5.825
S.D. OF GROUP SIZE IS    5.530

ALL MEANS AND S.D.S BASED ON UNIQUE TERM COUNT
```

Figure 10-32

PROFILE TERM STATISTICS

- mean frequency of LCB's in the terms
- standard deviation of term LCB frequency
- mean frequency of all possible LCB's
- mean group size (number of terms sorted under the average LCB)
- standard deviation of group size

10.4.5 Profile Term Frequency per Issue

The Profile Term Frequency per Issue list provides, for each term in each profile, the number of times that term appeared in the issue of the data base that was searched. (See Figure 10-33).

10.4.6 Profile Term and Hit, Cost Data-Summaries

Data and averages are generated for each production search of each issue of each data base regarding profiles, terms, citations and hits. The following statistics prepared for each issue are summarized on Tables 10-14 through 10-24.

- number of terms per profile
- aggregation ratio for profile terms
- total number of citations on the data base
- number of citations retrieved by profiles in the run
- average number citations retrieved (hits) per profile
- average number of profiles for which a retrieved citation was a hit
- average number of hits per profile normalized to the average number of citations per issue based on the complete volume
- computer cost per profile
- computer cost per profile averaged to date within the volume
- computer cost per term
- computer cost per term averaged to date within the volume
- computer cost per hit
- computer cost per hit averaged to date within the volume
- computer cost per profile term per citation
- computer cost per profile term per citation averaged to date within the volume

```
C1G090011C

 CASING                      1    FAULT*                   5    HYDRAUL*
 CAVIT*                     12    BRINE                    7    SUBSIDENCE
 DOME*                       4    DRILLING                11    TREE*
 DEPOSIT*                   89    DEPOSITION              20    FINITE ELEMENT
 FRACTUR*                   21    LOGGING                  3    STRENGTH
 POTASSIUM CHLORIDE*        19    SODIUM CHLORIDE         22    ROCK MECHANIC*
 TRONA                       1    MINING                   5    OIL
 SALT                       82    SALT BED*                0    SALT BEDS*
 STRATA*                     6    SOLUBL*                 18    UNCOUPL*
 WELL                       12    WELLS                    6    SOLUTION EXTRA


C1G100011B

 PAPER*                     73    STRENGTH*               98    BRIGHT*
 KRAFT                       2    PULP*                   24    SMOOTH*
 SURFACE                   147    RHEOLOGY                 2    PROPERT*
 ROUGH*                      2


C1L01C021A

 $CA043C00$                 90    PAPER*                  73    BINDER*
 CARDBOARD*                  0    ADHESIVE*               54    ELECTROCONDUCT
 ELECTROPHOTOGRAPH*          6    COLOR*                  70    COAT*
 SURFACE                   147    INTERNAL                18    CORRUGAT*
 WATER RESIST*               9    NEWSPRINT                0    SIZE*


C1LC10031A

 $CA035000$                241    $CA036000$             327    $CA037000$
 $CA041C00$                 11    $CA042000$             124    BLOCK
 MACEA                       0    ZIEGLER CATALYST*        0    GRAFT
*VINYL CHLORIDE             30   *ETHYLENE               231    JAPNA
 JPLPA                       0    JPYAA                    0    POLMA
 ANIONIC                     9    MAMOB                    0   *PROPYLENE
*STYRENE                   107
```

Figure 10-33
PROFILE TERM FREQUENCY/ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 7

	Terms		Cits.		Hits				Cost/Pr
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Norm./Prof. Even	Issue
9	14.0	21.8	3955	-	8.4	-	10.11		7.39
10	17.9	22.0	6249	-	16.5	-		11.53	10.22
11	17.8	22.0	4884	-	11.3	-	10.10		10.24
12	18.0	22.2	5958	-	12.7	-		9.26	11.32
13	16.5	21.3	5272	-	13.2	-	10.91		10.32
14	17.0	22.2	5465	-	11.1	-		8.83	11.50
15	23.0	20.2	3704	-	12.2	-	14.19		5.44
16	16.4	21.3	5245	1149*	10.6	1.30		8.99	4.33
17	14.3	19.4	4589	1785*	17.1	1.41	16.27		4.63
18	18.7	21.2	5697	1325*	12.3	1.24		9.39	4.40
19	21.6	19.7	4444	1741*	17.3	1.43	16.97		4.41
20	18.7	21.4	6246	1359*	12.0	1.18		8.39	5.10
21	17.1	20.9	4099	1528*	16.1	1.61	17.10		4.65
22	19.9	22.0	4287	1524*	12.3	1.16		12.48	6.24
23	23.2	21.1	4301	1531*	17.6	2.36	17.82		4.19

*Estimated

(Data for issue nos. 1-8
and 24-26 do not exist.)

Table 10-14
PROFILE TERM, HIT, COST DATA VS. ISSUE

MICAL ABSTRACTS CONDENSATES VOLUME 71

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per rofile	Hit/Ret. Cit.	Norm./Prof. Odd	Even	Issue	Avg.	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
8.4	-	10.11		7.39	7.39	.34	.34	.99	.89	9.44	9.44
16.5	-		11.53	10.22	8.81	.46	.40	.62	.76	7.42	8.43
11.3	-	10.10		10.24	9.28	.47	.42	.91	.81	9.53	8.80
12.7	-		9.26	11.32	9.79	.51	.44	.90	.83	8.58	8.74
13.2	-	10.91		10.32	9.90	.48	.45	.78	.82	9.19	8.83
11.1	-		8.83	11.50	10.17	.48	.45	.96	.84	8.73	8.82
12.2	-	14.19		5.44	9.49	.27	.43	.45	.79	7.28	8.60
10.6	1.30		8.99	4.33	8.85	.20	.40	.41	.74	3.82	8.00
17.1	1.41	16.27		4.63	8.37	.24	.38	.27	.69	5.19	7.69
12.3	1.24		9.39	4.40	7.98	.21	.37	.36	.66	3.65	7.28
17.3	1.43	16.97		4.41	7.65	.22	.35	.26	.62	5.02	7.08
12.0	1.18		8.39	5.10	7.44	.24	.34	.42	.60	3.79	6.80
16.1	1.61	17.10		4.65	7.23	.22	.33	.29	.58	5.41	6.70
12.3	1.16		12.48	6.24	7.16	.28	.33	.50	.57	6.62	6.69
17.6	2.36	17.82		4.19	6.96	.20	.32	.24	.55	4.60	6.55

(Data for issue nos. 1-8
and 24-26 do not exist.)

Table 10-14
FILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

	Terms		Cits.		Hits				Cost/Profile		C
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Norm./Prof. Even	Issue	Avg.	Is
1	23.6	21.2	3771	1462	15.7	1.62	18.90		3.91	3.91	
2	25.1	22.0	5296	1364	13.6	1.45		15.86	4.42	4.17	
3	24.9	21.4	3969	1757	18.3	1.67	20.92		3.69	4.01	
4	25.7	22.7	5254	1438	13.5	1.38		15.86	4.64	4.17	
5	24.9	21.6	3616	1451	14.2	1.54	17.83		3.38	4.01	
6	28.5	23.9	6310	2177	21.0	1.52		20.59	6.51	4.44	
7	27.6	21.9	3958	1689	16.4	1.61	18.85		3.78	4.33	
8	29.5	23.5	6468	1858	17.2	1.48		16.40	6.33	4.58	
9	27.6	21.7	5438	2203	21.9	1.61	18.33		5.01	4.62	
10	28.6	24.7	6629	1878	17.9	1.39		16.70	6.92	4.86	
11	26.9	23.6	5121	2088	21.7	1.56	19.26		5.42	4.91	
12	28.7	25.7	6650	2006	18.0	1.35		16.72	7.59	5.13	
13	26.5	23.9	4707	1903	21.7	1.72	20.97		4.98	5.12	
14	28.3	26.0	7291	2105	19.0	1.40		16.05	8.33	5.35	
15	26.7	23.9	4915	2012	21.3	1.66	19.66		5.27	5.34	
16	32.0	27.4	6572	1964	17.9	1.44		16.63	8.33	5.53	
17	26.2	24.1	4820	1995	20.7	1.63	19.47		5.02	5.50	
18	27.8	25.6	5604	1443	13.3	1.41		14.70	5.77	5.52	
19	26.3	23.8	4629	1832	17.9	1.56	17.54		4.66	5.47	
20	28.0	25.5	5743	1641	15.3	1.47		16.49	5.75	5.48	
21	26.8	26.4	4924	2117	25.1	1.61	23.13		3.89	5.40	
22	27.4	27.8	5644	1485	16.7	1.40		18.25	3.35	5.32	
23	26.8	26.7	4467	1818	23.0	1.59	23.45		2.87	5.21	
24	27.2	28.1	6549	1571	20.8	1.34		19.59	4.36	5.17	
25	26.3	27.4	4702	1858	25.4	1.56	24.48		2.80	5.08	
26	26.8	28.9	6287	1490	20.3	1.32		26.70	4.68	5.06	

Table 10-15

PROFILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per Profile	Hit/Ret. Cit.	Norm./Prof.								x 10^{-5}	
		Odd	Even	Issue	Avg.	Issue	Avg.	Issue	Avg.	Issue	Avg.
15.7	1.62	18.90		3.91	3.91	.18	.18	.25	.25	4.89	4.89
13.6	1.45		15.86	4.42	4.17	.20	.19	.33	.29	3.80	4.35
18.3	1.67	20.92		3.69	4.01	.17	.18	.20	.26	4.36	4.35
13.5	1.38		15.86	4.64	4.17	.20	.18	.34	.28	3.89	4.23
14.2	1.54	17.83		3.38	4.01	.16	.18	.24	.27	4.32	4.25
21.0	1.52		20.59	6.51	4.44	.27	.20	.31	.28	4.31	4.26
16.4	1.61	18.85		3.78	4.33	.17	.19	.23	.27	4.35	4.27
17.2	1.48		16.40	6.33	4.58	.27	.20	.37	.28	4.17	4.26
21.9	1.61	18.33		5.01	4.62	.23	.21	.23	.28	4.25	4.26
17.9	1.39		16.70	6.92	4.86	.28	.21	.39	.29	4.23	4.26
21.7	1.56	19.26		5.42	4.91	.23	.22	.25	.29	4.50	4.28
18.0	1.35		16.72	7.59	5.13	.30	.22	.42	.30	4.44	4.29
21.7	1.72	20.97		4.98	5.12	.21	.22	.23	.29	4.42	4.30
19.0	1.40		16.05	8.33	5.35	.32	.23	.44	.30	4.40	4.31
21.3	1.66	19.66		5.27	5.34	.22	.23	.25	.30	4.49	4.32
17.9	1.44		16.63	8.33	5.53	.30	.23	.47	.31	4.63	4.34
20.7	1.63	19.47		5.02	5.50	.21	.23	.24	.30	4.32	4.34
13.3	1.41		14.70	5.77	5.52	.23	.23	.43	.31	4.02	4.32
17.9	1.56	17.54		4.66	5.47	.20	.23	.26	.31	4.22	4.32
15.3	1.47		16.49	5.75	5.48	.23	.23	.37	.31	3.02	4.30
25.1	1.61	23.13		3.89	5.40	.15	.22	.16	.30	2.99	4.23
16.7	1.40		18.25	3.35	5.32	.12	.22	.20	.30	2.14	4.14
23.0	1.59	23.45		2.87	5.21	.11	.21	.12	.29	2.41	4.06
20.8	1.34		19.59	4.36	5.17	.16	.21	.21	.29	2.37	3.99
25.4	1.56	24.48		2.80	5.08	.10	.21	.11	.28	2.18	3.92
20.3	1.32		26.70	4.68	5.06	.16	.21	.23	.28	2.57	2.87

Table 10-15

PROFILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

	Terms		Cits.		Hits				Cost/Profile	
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Norm./Prof. Even	Issue	Avg
1	25.66	27.88	4182	1616	22.13	1.51	24.05		2.45	2.45
2	27.28	30.05	6321	1908	26.15	1.39		24.79	5.22	3.84
3	25.20	28.70	4731	2016	28.70	1.58	27.54		4.13	3.93
4	25.80	30.30	5855	1716	26.20	1.41		26.95	5.00	4.20
5	26.80	30.91	4734	1963	30.88	1.56	29.66		4.14	4.19
6	25.36	30.29	5523	1419	21.51	1.38		23.34	3.82	4.14
7	26.02	32.32	4351	1740	28.14	1.50	29.41		3.80	4.08
8	26.32	31.26	5887	1438	22.75	1.38		23.16	3.90	4.06
9	25.36	32.47	4190	1705	27.16	1.50	29.12		3.43	3.99
10	26.54	32.22	6275	1502	23.84	1.38		22.77	4.08	4.00
11	31.66	33.50	4306	1862	31.83	1.64	33.61		3.54	3.96
12	28.70	32.47	6016	1493	23.08	1.41		22.99	3.84	3.95
13	30.78	33.74	4425	1874	33.48	1.63	34.40		4.75	4.04
14	27.04	33.01	5974	1594	28.35	1.44		28.43	4.29	4.06
15	31.59	34.21	4737	2019	35.25	1.61	38.30		3.82	4.05
16	27.14	33.63	6056	1963	38.58	1.55		38.17	4.88	4.09
17	21.53	36.05	4487	2191	40.74	1.56	41.28		4.61	4.12
18	22.83	33.99	6253	1994	38.04	1.51		36.46	4.78	4.16
19	21.93	35.16	4774	2175	36.50	1.44	34.77		4.76	4.19
20	23.41	32.48	6039	1901	31.19	1.44		30.95	4.61	4.21
21	22.07	34.53	4880	2175	35.68	1.48	33.24		5.06	4.23
22	24.40	33.70	6042	1824	31.13	1.48		30.89	4.46	4.24
23	22.70	33.30	4717	2083	32.58	1.51	31.41		4.48	4.25
24	24.50	32.13	5800	1676	26.35	1.47		27.22	4.07	4.24
25	22.70	33.46	4595	2143	33.98	1.53	33.64		4.37	4.25
26	24.50	32.27	5862	1710	25.89	1.42		26.47	4.23	4.25

Table 10-16

PROFILE TERM, HIT. COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Even	Issue	Avg	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
22.13	1.51	24.05		2.45	2.45	.09	.13	.11	.16	2.10	2.10
26.15	1.39		24.79	5.22	3.84	.17	.13	.20	.16	2.68	2.39
28.70	1.58	27.54		4.13	3.93	.14	.13	.14	.15	2.95	2.58
26.20	1.41		26.95	5.00	4.20	.17	.14	.19	.16	2.81	2.63
30.88	1.56	29.66		4.14	4.19	.13	.14	.13	.15	2.83	2.67
21.51	1.38		23.34	3.82	4.14	.13	.14	.18	.16	2.29	2.61
28.14	1.50	29.41		3.80	4.08	.12	.14	.14	.16	2.71	2.62
22.75	1.38		23.16	3.90	4.06	.12	.13	.17	.16	2.12	2.56
27.16	1.50	29.12		3.43	3.99	.11	.13	.13	.15	2.52	2.55
23.84	1.38		22.77	4.08	4.00	.13	.13	.17	.16	2.02	2.50
31.83	1.64	33.61		3.54	3.96	.11	.13	.11	.15	2.46	2.50
23.08	1.41		22.99	3.84	3.95	.12	.13	.17	.15	1.96	2.45
33.48	1.63	34.40		4.75	4.04	.14	.13	.14	.15	3.09	2.50
28.35	1.44		28.43	4.29	4.06	.13	.12	.15	.15	2.18	2.48
35.25	1.61	38.30		3.82	4.05	.11	.13	.11	.15	2.36	2.47
38.58	1.55		38.17	4.88	4.09	.15	.13	.13	.15	2.40	2.47
40.74	1.56	41.28		4.61	4.12	.13	.13	.11	.15	2.85	2.49
38.04	1.51		36.46	4.78	4.16	.14	.13	.13	.15	2.25	2.48
36.50	1.44	34.77		4.76	4.19	.14	.13	.13	.14	2.83	2.49
31.19	1.44		30.95	4.61	4.21	.14	.13	.15	.14	2.35	2.49
35.68	1.48	33.24		5.06	4.23	.15	.13	.14	.14	3.00	2.51
31.13	1.48		30.89	4.46	4.24	.13	.13	.14	.14	2.19	2.50
32.58	1.51	31.41		4.48	4.25	.13	.13	.14	.14	2.85	2.51
26.35	1.47		27.22	4.07	4.24	.13	.13	.15	.14	2.18	2.50
33.98	1.53	33.64		4.37	4.25	.13	.13	.13	.14	2.84	2.51
25.89	1.42		26.47	4.23	4.25	.13	.13	.16	.14	2.24	2.48

Table 10-16

PROFILE TERM, HIT. COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

	Terms		Cits.		Hits				Cost/Profile	
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Norm./Prof.			
							Odd	Even	Issue	Avg.
1	22.5	32.4	3858	1690	24.7	1.45	32.04		3.69	3.69
2	24.1	30.9	5577	1759	26.2	1.43		31.09	4.11	3.90
3	22.5	32.4	3961	1760	26.3	1.48	33.12		3.70	3.83
4	24.1	29.8	5738	1735	23.2	1.42		26.82	3.94	3.86
5	22.7	31.4	3957	1761	22.7	1.43	28.67		3.81	3.85
6	24.9	30.2	4746	1409	20.1	1.45		28.09	3.39	3.77
7	23.2	30.0	4087	1895	24.9	1.47	30.39		3.25	3.70
8	28.9	29.1	6816	2435	31.5	1.45		36.99	4.18	3.76
9	23.4	30.0	4475	2005	26.0	1.47	29.04		4.00	3.79
10	25.5	28.8	5986	2090	28.1	1.45		27.31	4.48	3.86
11	25.1	29.3	5618	2262	30.1	1.56	31.65		3.86	3.86
12	28.2	29.1	6553	2044	28.8	1.62		29.12	4.29	3.90
13	25.3	28.4	3667	1930	25.3	1.65	34.40		4.52	3.95
14	28.2	28.1	7074	2540	32.4	1.67		30.32	4.77	4.01
15	25.2	28.4	5174	2705	34.7	1.61	44.29		6.31	4.16
16	28.5	27.9	6190	1983	23.6	1.57		25.25	4.52	4.18
17	24.1	28.2	5458	1812	26.9	1.50	29.22		5.27	4.24
18	25.1	27.9	6350	1855	21.7	1.48		16.57	4.18	4.24
19	24.3	27.9	6243	3026	38.0	1.60	30.30		5.37	4.30
20	25.4	27.6	6519	2023	21.5	1.43		15.22	4.21	4.29
21	23.9	27.8	5458	2752	34.5	1.59	31.54		4.93	4.32
22	25.3	27.8	7442	2565	28.7	1.48		25.53	4.85	4.34
23	25.5	28.2	6151	2842	34.9	1.53	28.32		4.89	4.36
24	27.4	29.2	8732	3154	38.6	1.58		29.30	5.96	4.43
25	25.8	28.3	7531	3543	43.5	1.55	28.83		5.97	4.49
26	28.9	28.7	5722	2837	32.7	1.60		25.10	5.28	4.52

Table 10-17

PROFILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per Profile	Hit/Ret. Cit.	Norm./Prof.								x 10^{-5}	
		Odd	Even	Issue	Avg.	Issue	Avg.	Issue	Avg.	Issue	Avg.
24.7	1.45	32.04		3.69	3.69	.11	.11	.15	.15	2.96	2.96
26.2	1.43		31.09	4.11	3.90	.13	.12	.16	.16	2.38	2.67
26.3	1.48	33.12		3.70	3.83	.11	.12	.14	.15	2.88	2.74
23.2	1.42		26.82	3.94	3.86	.13	.12	.17	.16	2.30	2.63
22.7	1.43	28.67		3.81	3.85	.12	.12	.17	.16	3.07	2.72
20.1	1.45		28.09	3.39	3.77	.11	.12	.17	.16	2.36	2.66
24.9	1.47	30.39		3.25	3.70	.11	.12	.13	.16	2.65	2.66
31.5	1.45		36.99	4.18	3.76	.14	.12	.12	.15	2.33	2.62
26.0	1.47	29.04		4.00	3.79	.13	.12	.15	.15	2.98	2.66
28.1	1.45		27.31	4.48	3.86	.16	.13	.16	.15	2.29	2.62
30.1	1.56	31.65		3.86	3.86	.13	.13	.13	.15	2.78	2.63
28.8	1.62		29.12	4.29	3.90	.15	.13	.15	.15	2.25	2.60
25.3	1.65	34.40		4.52	3.95	.16	.13	.18	.15	4.73	2.77
32.4	1.67		30.32	4.77	4.01	.17	.13	.15	.15	2.39	2.74
34.7	1.61	44.29		6.31	4.16	.22	.14	.18	.15	4.28	2.84
23.6	1.57		25.25	4.52	4.18	.16	.14	.19	.16	2.61	2.83
26.9	1.50	29.22		5.27	4.24	.19	.14	.16	.16	3.43	2.86
21.7	1.48		16.57	4.18	4.24	.15	.14	.29	.16	2.56	2.85
38.0	1.60	30.30		5.37	4.30	.19	.15	.14	.16	3.08	2.86
21.5	1.43		15.22	4.21	4.29	.15	.15	.28	.17	2.33	2.83
34.5	1.59	31.54		4.93	4.32	.18	.15	.14	.17	3.24	2.85
28.7	1.48		25.53	4.85	4.34	.17	.15	.17	.17	2.34	2.83
34.9	1.53	28.32		4.89	4.36	.17	.15	.14	.17	2.82	2.83
38.6	1.58		29.30	5.96	4.43	.20	.15	.15	.17	2.33	2.81
43.5	1.55	28.83		5.97	4.49	.21	.15	.14	.16	2.80	2.81
32.7	1.60		25.10	5.28	4.52	.18	.16	.16	.16	2.14	2.78

Table 10-17

PROFILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

Issue	Terms Agg. Ratio	Terms Per Profile	Cits. Total	Cits. Ret.	Hits Per Profile	Hits Hit/Ret. Cit.	Hits Norm./Prof. Odd	Hits Norm./Prof. Even	Cost/Profil Issue	Cost/Profil Avg
1	24.9	27.0	6342	3055	37.7	1.65	30.81		4.33	4.3
2	22.8	25.9	8363	2846	30.8	1.45		25.19	4.00	4.1
3	24.9	27.0	6260	3186	39.8	1.67	32.93		4.38	4.2
4	22.8	25.9	8720	3271	38.1	1.56		29.87	4.36	4.2
5	21.6	26.5	6917	3254	43.0	1.58	32.15		4.89	4.3
6	22.6	25.0	7964	2847	31.9	1.52		27.39	3.95	4.3
7	21.4	25.6	4906	2367	29.3	1.57	30.91		3.50	4.2
8	22.3	24.4	8856	3099	32.5	1.48		25.10	4.18	4.2
9	22.0	25.3	5803	2689	31.5	1.57	28.07		4.02	4.1
10	22.6	24.2	6870	2358	24.8	1.52		24.63	3.26	4.0
11	21.9	25.2	6510	2864	33.7	1.58	28.42		4.01	4.0
12	21.6	24.0	6694	2320	26.0	1.52		26.52	3.17	4.0
13	20.9	24.0	4960	2309	26.8	1.48	27.95		2.60	3.9
14	21.8	23.5	5461	1789	20.1	1.54		25.20	1.99	3.7
15	21.8	23.5	4362	1551	14.9	1.32	17.76		2.20	3.6
16	21.7	23.8	6048	1985	22.1	1.50		24.97	2.74	3.6
17	20.9	24.1	5690	2597	29.2	1.40	26.56		2.96	3.5
18	21.6	23.8	4989	1685	19.5	1.53		26.75	2.34	3.4
19	22.1	24.3	3016	1292	15.3	1.45	26.39		1.64	3.4
20	22.1	23.5	7438	2692	30.3	1.55		27.81	3.18	3.3
21	21.0	25.0	4030	1637	20.9	1.44	26.81		2.18	3.3
22	21.0	24.1	5681	2548	30.8	1.45		31.48	3.11	3.32
23	20.4	25.0	4328	1737	22.5	1.49	26.95		2.40	3.2
24	20.0	26.4	4120	2551	32.5	1.44		53.81	3.22	3.2
25	20.4	24.8	4481	1807	22.8	1.45	26.32		2.54	3.2
26	20.1	24.6	4159	2369	32.5	1.58		53.40	3.01	3.24

Table 10-18

PROFILE TERM, HIT, COST DATA VS. ISSUE

EMICAL ABSTRACTS CONDENSATES VOLUME 75

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Even	Issue	Avg.	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
37.7	1.65	30.81		4.33	4.33	.16	.16	.11	.11	2.53	2.53
30.8	1.45		25.19	4.00	4.17	.15	.15	.13	.12	1.84	2.19
39.8	1.67	32.93		4.38	4.24	.16	.16	.11	.12	2.59	2.32
38.1	1.56		29.87	4.36	4.27	.17	.16	.11	.12	1.93	2.22
43.0	1.58	32.15		4.89	4.39	.18	.16	.11	.12	2.66	2.31
31.9	1.52		27.39	3.95	4.32	.16	.16	.12	.12	1.98	2.26
29.3	1.57	30.91		3.50	4.20	.14	.16	.12	.12	2.78	2.33
32.5	1.48		25.10	4.18	4.20	.17	.16	.13	.12	1.93	2.28
31.5	1.57	28.07		4.02	4.18	.16	.16	.13	.12	2.74	2.33
24.8	1.52		24.63	3.26	4.09	.13	.16	.13	.12	1.96	2.29
33.7	1.58	28.42		4.01	4.08	.16	.16	.12	.12	2.60	2.32
26.0	1.52		26.52	3.17	4.00	.13	.16	.12	.12	1.96	2.29
26.8	1.48	27.95		2.60	3.90	.1	15	.10	.12	2.18	2.28
20.1	1.54		25.20	1.99	3.76	.08	.15	.10	.12	1.54	2.23
14.9	1.32	17.76		2.20	3.66	.09	.14	.11	.12	2.00	2.21
22.1	1.50		24.97	2.74	3.60	.11	.14	.12	.12	1.90	2.20
29.2	1.40	26.56		2.96	3.56	.12	.14	.10	.12	2.15	2.19
19.5	1.53		26.75	2.34	3.49	.10	.14	.12	.12	1.96	2.18
15.3	1.45	26.39		1.64	3.40	.07	.13	.11	.12	2.23	2.18
30.3	1.55		27.81	3.18	3.39	.14	.13	.10	.12	1.81	2.16
20.9	1.44	26.81		2.18	3.33	.09	.13	.10	.11	2.15	2.16
30.8	1.45		31.48	3.11	3.32	.13	.13	.10	.11	1.92	2.15
22.5	1.49	26.95		2.40	3.28	.10	.13	.11	.11	2.21	2.15
32.5	1.44		53.81	3.22	3.28	.13	.13	.10	.11	2.07	2.15
22.8	1.45	26.32		2.54	3.25	.10	.13	.11	.11	2.27	2.16
32.5	1.58		53.40	3.01	3.24	.12	.13	.16	.11	2.10	2.15

Table 10-18

PROFILE TERM, HIT, COST DATA VS. ISSUE

	Terms		Cits.		Hits				Cost/Profil	
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Norm./Prof. Even	Issue	Avg
1	20.1	24.6	4159	1717	22.3	1.49	28.34		1.92	1.9
2	23.1	23.6	4816	2441	30.7	1.77		44.77	1.81	1.8
3	23.0	24.0	6051	1901	20.8	1.43	18.18		2.07	1.9
4	23.2	23.3	6051	2679	30.9	1.79		35.92	2.36	2.0
5	23.1	23.6	4592	1790	18.8	1.39	21.60		1.94	2.0
6	24.2	23.1	4126	2180	24.1	1.78		41.11	1.83	2.0
7	25.0	24.0	7187	1813	19.9	1.52	25.41		1.76	1.9
8	24.9	23.8	7187	3033	33.9	1.76		33.15	2.87	2.0
9	25.3	24.8	4732	2229	25.7	1.57	28.70		2.24	2.1
10	28.1	24.6	7625	2394	42.3	2.69		38.98	3.06	2.2
11	26.7	23.9	5923	2874	30.9	1.58	27.50		2.46	2.2
12	27.8	23.9	7887	3139	40.2	1.89		35.80	2.87	2.2
13	26.5	23.5	4806	2739	32.2	1.72	35.32		1.92	2.2
14	27.6	23.7	8103	3487	46.3	1.91		40.19	3.04	2.3
15	28.5	24.0	5236	2609	26.9	1.56	27.15		2.14	2.2
16	28.4	23.3	8200	3747	47.9	1.92		41.07	3.05	2.3
17	27.6	23.4	5706	3052	32.3	1.65	29.84		2.12	2.3
18	29.2	23.1	7620	3335	40.9	1.97		37.76	2.08	2.3
19	27.7	24.0	5815	2874	31.0	1.65	28.10		1.79	2.2
20	28.7	24.2	5815	3691	49.7	2.04		45.36	2.22	2.2
21	27.9	24.6	6153	3053	34.3	1.66	28.36		1.69	2.2
22	28.3	25.3	7900	3678	47.7	1.93		42.46	2.20	2.2
23	25.8	25.4	5910	2951	34.0	1.61	30.38		1.69	2.2
24	27.8	25.9	7258	3374	47.2	2.00		45.68	2.23	2.2
25	25.0	26.8	5541	2819	21.8	1.64	34.19		1.67	2.2
26	27.8	26.9	7870	4026	51.2	2.01		51.90	2.50	2.2

Table 10-19

PROFILE TERM, HIT, COST DATA VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

Hits				Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Per Profile	Hit/Ret. Cit.	Norm./Prof. Odd	Norm./Prof. Even	Issue	Avg.	Issue	Avg.	Issue	Avg.	Issue	Avg.
22.3	1.49	28.34		1.92	1.92	.08	.08	.09	.09	1.87	1.87
30.7	1.77		44.77	1.81	1.87	.08	.08	.06	.08	1.29	1.58
20.8	1.43	18.18		2.07	1.94	.09	.08	.10	.08	1.79	1.65
30.9	1.79		35.92	2.36	2.05	.10	.09	.07	.08	1.63	1.65
18.8	1.39	21.60		1.94	2.03	.08	.09	.10	.09	1.78	1.67
24.1	1.78		41.11	1.83	2.00	.08	.09	.08	.08	1.63	1.67
19.9	1.52	25.41		1.76	1.97	.07	.08	.09	.08	1.77	1.68
33.9	1.76		33.15	2.87	2.08	.12	.09	.10	.09	1.67	1.68
25.7	1.57	28.70		2.24	2.10	.09	.09	.09	.09	1.91	1.70
42.3	2.69		38.98	3.06	2.20	.12	.09	.10	.09	1.62	1.70
30.9	1.58	27.50		2.46	2.22	.10	.09	.08	.09	1.73	1.70
40.2	1.89		35.80	2.87	2.27	.12	.09	.07	.09	1.52	1.68
32.2	1.72	35.32		1.92	2.24	.08	.09	.06	.08	1.69	1.69
46.3	1.91		40.19	3.04	2.30	.13	.10	.13	.09	1.58	1.68
26.9	1.56	27.15		2.14	2.29	.09	.10	.08	.09	1.70	1.68
47.9	1.92		41.07	3.05	2.34	.13	.10	.06	.09	1.59	1.67
32.3	1.65	29.84		2.12	2.33	.09	.10	.07	.08	1.59	1.67
40.9	1.97		37.76	2.08	2.32	.09	.10	.05	.08	1.18	1.64
31.0	1.65	28.10		1.79	2.29	.07	.10	.06	.08	1.28	1.62
49.7	2.04		45.36	2.22	2.29	.09	.10	.04	.08	1.18	1.60
34.3	1.66	28.36		1.69	2.26	.07	.09	.05	.08	1.11	1.58
47.7	1.93		42.46	2.20	2.26	.09	.09	.05	.08	1.10	1.56
34.0	1.61	30.38		1.69	2.24	.07	.09	.05	.08	1.13	1.54
47.2	2.00		45.68	2.23	2.24	.09	.09	.05	.08	1.18	1.53
21.8	1.64	34.19		1.67	2.22	.06	.09	.05	.07	1.12	1.51
51.2	2.01		51.90	2.50	2.23	.09	.09	.04	.07	1.17	1.50

Table 10-19

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIORESEARCH INDEX VOLUME 71

	Terms		Cits.		Hits		Cost/Profile		Cost/Term	
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue	Avg.
1	13.30	29.4	7500	1410	33.5	1.22	9.43	9.43	.32	.32
2	11.60	25.7	7500	890	24.7	1.11	11.83	10.63	.46	.39
3	11.50	25.6	7500	731	20.3	1.07	9.29	10.18	.36	.38
4	12.20	24.2	5833	1241	28.2	1.11	7.19	9.44	.30	.36
5	12.40	23.3	7500	1393	29.6	1.15	7.68	9.08	.33	.35
6	12.70	22.9	7500	1542	30.8	1.13	7.65	8.85	.33	.35
7	12.20	19.3	7500	4788	77.2	1.46	8.44	8.79	.40	.36
8	12.10	18.8	7500	2625	40.3	1.26	5.76	8.41	.31	.35
9	10.30	19.3	7500	2850	46.7	1.25	6.56	8.20	.34	.35
10	14.00	18.8	7500	2611	38.3	1.23	5.04	7.89	.28	.34
11	11.80	20.1	7500	2130	40.1	1.19	7.87	7.89	.39	.35
12	11.70	20.4	7500	2501	48.0	1.24	7.71	7.87	.38	.35

TABLE 10-20

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIORESEARCH INDEX VOLUME 71

.	Hits		Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
1410	33.5	1.22	9.43	9.43	.32	.32	.28	.28	4.28	4.28
890	24.7	1.11	11.83	10.63	.46	.39	.48	.38	6.12	5.20
731	20.3	1.07	9.29	10.18	.36	.38	.46	.41	4.83	5.08
1241	28.2	1.11	7.19	9.44	.30	.36	.28	.38	3.96	4.80
1393	29.6	1.15	7.68	9.08	.33	.35	.26	.35	4.38	4.71
1542	30.8	1.13	7.65	8.85	.33	.35	.25	.34	4.45	4.67
4788	77.2	1.46	8.44	8.79	.40	.36	.12	.30	5.34	4.77
2625	40.3	1.26	5.76	8.41	.31	.35	.14	.28	4.08	4.68
2850	46.7	1.25	6.56	8.20	.34	.35	.14	.27	4.52	4.66
2611	38.3	1.23	5.04	7.89	.28	.34	.13	.25	6.89	4.89
2130	40.1	1.19	7.87	7.89	.39	.35	.20	.25	5.20	4.91
2501	48.0	1.24	7.71	7.87	.38	.35	.16	.24	5.03	4.92

TABLE 10-20

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

	Terms		Cits.		Hits		Cost/Profile		Cost/Te
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg	Issue
(Data for Issues 1-6 do not exist.)									
7	17.31	38.59	5832	681	30.95	1.15	6.38	6.38	.17
8	17.31	38.59	5833	628	28.55	1.15	5.93	6.16	.15
9	17.31	38.59	4088	510	28.18	1.18	4.52	5.61	.12
10	16.44	41.00	5833	967	42.04	1.16	10.27	6.78	.25
11	16.25	34.79	5833	945	32.59	1.25	7.73	6.97	.22
12	15.42	33.15	5835	1023	37.89	1.22	8.07	7.15	.24
13	15.42	33.15	5833	845	31.30	1.19	7.37	7.18	.22
14	18.06	32.66	5834	836	28.83	1.19	6.60	7.11	.20
15	18.31	33.13	5674	986	32.87	1.26	6.66	7.06	.20
16	18.29	33.17	5833	990	33.00	1.27	5.72	6.93	.17
17	16.13	34.00	5836	1112	35.87	1.18	5.92	6.83	.17
18	16.22	33.81	5837	975	31.45	1.28	5.78	6.73	.17
19	16.55	31.77	5836	1133	32.37	1.19	7.80	6.82	.25
20	16.84	31.23	5838	1053	30.09	1.22	7.15	6.85	.23
21	14.35	32.11	5836	2585	68.03	1.37	6.61	6.83	.21
22	13.00	31.46	5836	2192	59.24	1.28	7.19	6.86	.23
23	15.64	29.80	5834	1371	29.80	1.26	6.26	6.82	.21
24	15.64	29.41	5833	1357	29.41	1.25	5.59	6.75	.19

Table 10-21

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

s.	Hits		Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
(Data for Issues 1-6 do not exist.)										
681	30.95	1.15	6.38	6.38	.17	.17	.21	.21	2.83	2.83
628	28.55	1.15	5.93	6.16	.15	.16	.21	.21	2.64	2.74
510	28.18	1.18	4.52	5.61	.12	.15	.20	.21	2.87	2.78
967	42.04	1.16	10.27	6.78	.25	.17	.24	.22	4.30	3.16
945	32.59	1.25	7.73	6.97	.22	.18	.24	.22	3.81	3.29
1023	37.89	1.22	8.07	7.15	.24	.19	.21	.22	4.17	3.44
845	31.30	1.19	7.37	7.18	.22	.20	.24	.22	3.81	3.49
836	28.83	1.19	6.60	7.11	.20	.20	.23	.22	3.47	3.49
986	32.87	1.26	6.66	7.06	.20	.20	.20	.22	3.54	3.49
990	33.00	1.27	5.72	6.93	.17	.19	.17	.22	2.96	3.44
1112	35.87	1.18	5.92	6.83	.17	.19	.16	.21	2.98	3.40
975	31.45	1.28	5.78	6.73	.17	.19	.18	.21	2.93	3.36
1133	32.37	1.19	7.80	6.82	.25	.20	.24	.21	4.21	3.42
1053	30.09	1.22	7.15	6.85	.23	.20	.24	.21	3.92	3.46
2585	68.03	1.37	6.61	6.83	.21	.20	.10	.21	3.53	3.46
2192	59.24	1.28	7.19	6.86	.23	.20	.12	.20	3.91	3.49
1371	29.80	1.26	6.26	6.82	.21	.20	.21	.20	3.63	3.50
1357	29.41	1.25	5.59	6.75	.19	.20	.19	.20	3.24	3.49

Table 10-21

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

	Terms		Cits.		Hits		Cost/Profile		Cost/
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue
1	15.70	29.6	5833	1675	36.4	1.28	6.20	6.20	.21
2	15.10	31.0	5833	1671	37.1	1.28	7.54	6.87	.24
3	13.70	29.4	5834	1413	35.3	1.23	7.19	6.98	.24
4	13.60	29.4	5833	1411	34.4	1.21	7.80	7.18	.26
5	11.60	25.7	5833	754	20.9	1.14	7.30	7.21	.28
6	11.60	25.7	5834	801	22.2	1.12	7.34	7.23	.28
7	11.50	25.6	5833	717	19.9	1.08	7.64	7.29	.30
8	11.50	25.6	5833	717	19.9	1.08	7.55	7.32	.29
9	11.81	25.1	5833	789	19.5	1.12	7.03	7.29	.28
10	11.81	25.1	5833	913	22.9	1.13	7.76	7.34	.31
11	12.70	22.9	5834	1693	33.8	1.21	7.30	7.33	.32
12	15.10	21.0	5834	3155	46.3	1.43	5.93	7.22	.28
13	15.10	21.0	5833	2937	43.1	1.42	5.00	7.04	.24
14	15.10	21.0	5833	2805	41.2	1.38	5.71	6.95	.27
15	11.23	18.6	5833	3552	52.7	1.48	3.16	6.70	.17
16	11.50	18.6	5836	2142	33.4	1.28	5.81	6.64	.31
17	10.80	18.9	5832	2184	36.4	1.27	5.53	6.58	.29
18	10.80	18.9	5836	2260	37.6	1.29	6.05	6.55	.32
19	10.80	18.9	5836	2310	38.5	1.29	3.99	6.41	.21
20	10.80	18.9	5833	2241	37.3	1.27	5.98	6.39	.32
21	11.30	19.4	5839	2463	37.8	1.34	5.53	6.35	.29
22	11.60	19.3	5833	2447	38.8	1.31	5.77	6.32	.30
23	11.60	19.3	5833	2168	34.4	1.30	5.35	6.28	.28
24	11.70	20.4	5836	1830	35.1	1.21	6.57	6.19	.32

Table 10-22

PROFILE TERM, HIT, COST DATA VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

its.	Hits		Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
1675	36.4	1.28	6.20	6.20	.21	.21	.17	.17	3.59	3.59
1671	37.1	1.28	7.54	6.87	.24	.23	.20	.19	4.16	3.88
1413	35.3	1.23	7.19	6.98	.24	.23	.20	.19	4.19	3.98
1411	34.4	1.21	7.80	7.18	.26	.24	.23	.20	4.53	4.12
754	20.9	1.14	7.30	7.21	.28	.25	.35	.23	4.86	4.27
801	22.2	1.12	7.34	7.23	.28	.25	.33	.25	4.88	4.37
717	19.9	1.08	7.64	7.29	.30	.26	.38	.27	5.10	4.47
717	19.9	1.08	7.55	7.32	.29	.26	.38	.28	5.04	4.54
789	19.5	1.12	7.03	7.29	.28	.26	.36	.29	4.78	4.57
913	22.9	1.13	7.76	7.34	.31	.27	.34	.29	5.28	4.64
1693	33.8	1.21	7.30	7.33	.32	.27	.22	.29	5.46	4.72
3155	46.3	1.43	5.93	7.22	.28	.27	.22	.28	4.83	4.72
2937	43.1	1.42	5.00	7.04	.24	.27	.12	.27	4.07	4.67
2805	41.2	1.38	5.71	6.95	.27	.27	.14	.26	4.65	4.67
3552	52.7	1.48	3.16	6.70	.17	.27	.06	.25	2.89	4.55
2142	33.4	1.28	5.81	6.64	.31	.27	.17	.24	5.34	4.60
2184	36.4	1.27	5.53	6.58	.29	.27	.15	.24	5.00	4.63
2260	37.6	1.29	6.05	6.55	.32	.27	.16	.23	5.46	4.67
2310	38.5	1.29	3.99	6.41	.21	.27	.10	.23	3.60	4.62
2241	37.3	1.27	5.98	6.39	.32	.27	.16	.22	5.41	4.66
2463	37.8	1.34	5.53	6.35	.29	.27	.15	.22	4.91	4.67
2447	38.8	1.31	5.77	6.32	.30	.27	.15	.22	5.10	4.69
2168	34.4	1.30	5.35	6.28	.28	.27	.16	.21	4.72	4.69
1830	35.1	1.21	6.57	6.19	.32	.28	.19	.21	5.50	4.72

Table 10-22

PROFILE TERM, HIT, COST DATA VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUME 71

	Terms		Cits.		Hits		Cost/Profile		Cost
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue
1	10.0	20.0	5743	270	20.7	1.03	4.12	4.12	.21
2	11.2	19.2	5600	848	59.4	1.15	8.91	6.52	.46
3	16.0	16.7	5743	1183	56.3	1.54	3.62	5.55	.22
(Data for Issues 4-5 do not exist.)									
6	16.0	16.7	5743	1471	70.0	1.51	3.97	5.16	.24
7	19.1	15.5	5743	1858	71.4	1.47	3.53	4.83	.23
8	17.2	16.4	5743	1819	62.7	1.41	3.28	4.57	.20
9	17.2	16.4	7710	2131	73.4	1.44	3.54	4.42	.22
10	21.9	16.3	7116	3776	104.8	2.30	3.47	4.31	.21
11	22.1	14.5	7157	2179	60.5	1.51	2.82	4.14	.19
12	23.5	15.6	8320	2677	68.6	1.50	3.55	4.08	.15

Table 10-23

PROFILE TERM, HIT, COST DATA VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUME 71

its.	Hits		Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit.	
Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue	Avg.	Issue	Avg.	x 10^{-5} Issue	Avg.
270	20.7	1.03	4.12	4.12	.21	.21	.20	.20	3.82	3.82
848	59.4	1.15	8.91	6.52	.46	.34	.15	.18	8.20	6.01
1183	56.3	1.54	3.62	5.55	.22	.30	.06	.14	3.77	5.26
(Data for Issues 4-5 do not exist.)										
1471	70.0	1.51	3.97	5.16	.24	.28	.06	.12	3.50	4.82
1858	71.4	1.47	3.53	4.83	.23	.27	.05	.10	3.30	3.52
1819	62.7	1.41	3.28	4.57	.20	.26	.05	.10	2.93	4.25
2131	73.4	1.44	3.54	4.42	.22	.25	.05	.09	2.80	4.05
3776	104.8	2.30	3.47	4.31	.21	.25	.03	.08	2.98	3.91
2179	60.5	1.51	2.82	4.14	.19	.24	.05	.08	2.70	3.77
2677	68.6	1.50	3.55	4.08	.15	.23	.07	.08	1.82	3.58

Table 10-23

PROFILE TERM, HIT, COST DATA VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUME 72

	Terms		Cits.		Hits		Cost/Profile	
Issue	Agg. Ratio	Per Profile	Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.
1	25.3	20.3	6126	2127	33.2	1.44	2.35	2.35
2	30.0	20.7	4385	2206	27.2	1.67	1.84	2.10
3	27.1	20.0	5129	2719	36.2	1.49	1.95	2.05
4	29.1	19.9	5823	3229	41.3	1.63	2.24	2.10
5	29.0	20.2	5815	3665	48.8	1.57	2.80	2.24

(Data for Issues 6-26 do not exi

Table 10-24

PROFILE TERM, HIT, COST DATA VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUME 72

Cits.		Hits		Cost/Profile		Cost/Term		Cost/Hit		Cost/Term/Cit. $\times 10^{-5}$	
Total	Ret.	Per Profile	Hit/Ret. Cit.	Issue	Avg.	Issue	Avg.	Issue	Avg.	Issue	Avg.
6126	2127	33.2	1.44	2.35	2.35	.12	.12	.07	.07	1.88	1.88
4385	2206	27.2	1.67	1.84	2.10	.09	.11	.07	.07	2.02	1.95
5129	2719	36.2	1.49	1.95	2.05	.10	.10	.05	.06	1.90	1.93
5823	3229	41.3	1.63	2.24	2.10	.11	.11	.05	.06	1.93	1.93
5815	3665	48.8	1.57	2.80	2.24	.14	.11	.06	.06	1.86	1.92

(Data for Issues 6-26 do not exist.)

Table 10-24

PROFILE TERM, HIT, COST DATA VS. ISSUE

Another check we made for several issues of CA was the number of hits generated from each section of CA. This is not a regularly prepared data item but was done once to determine whether there were any sections that did not prove fruitful for our users. We thought we might be able to eliminate such sections from the search and thereby reduce cost--assuming the nature of the user group would not change in the area of the eliminated sections. We found that our users got hits from every section of CA. This is due to the fact that CSC has a very heterogeneous group of users.

To summarize, while we at IITRI are providing retrieval in the very practical, almost business-oriented, mode, we are not merely feeding profiles and data bases into a matching machine. We are doing considerable research regarding the entire operational system. More data on this are given in Section 11.

Without a totally controlled system in which vocabularies, data base formats, and record contents and formats are controlled and without a static software system, compiler, hardware configuration and operating system etc., there is no reasonable way to maintain an overview without maintaining and interpreting such data to guide future efforts.

10.4.7 Personnel

The personnel tasks involved in design and operation of a center include:

- management
- system design
- programming --develop and maintain software system including adapting to data base changes
- profile coordinating and user liaison
- keypunching --programs and profiles
- clerical tasks --maintain records and distribute output
- promotion and marketing
- tape library maintenance

CSC maintains weekly records of man hours per week per function in order to monitor current expenditures, monitor staff performance, determine profile costs and estimate future rates.

While data base leases and royalties, machine time, travel, purchased materials and postage are significant budget items the major expenditure in a center is personnel salaries.

10.5 Marketing

10.5.1 Mailings

The Computer Search Center has used direct mail campaigns to acquaint large numbers of people with the CSC's services, as well as to inform selected groups of people about specific activities. For example, approximately 5000 brochures sent to announce a CSC workshop on "Computer Retrieval of Scientific Information" serve not only to solicit the 20 or 30 workshop attendees but to help keep the CSC associated in people's minds with information retrieval. Mass mailings serve a publicity function rather than as a mechanism for directly soliciting SDI subscribers. The dates, number of items sent, recipients, and responses are listed for CSC direct mailings in the following list.

Date	No. Items Sent/Responses	Recipients
July 1970	800/1	Presidents of chemical companies with over 1000 employees
September 1970	approx. 2000 + 560 /95	IEEE subscribers IEEE midwestern members
September 1970	135/22	Members of ACS Div. of Chem. Lit., Chicago Sec.
November 1970	275/22	Major U.S. universities
November 1970	approx. 5000/19	ASIS Members and previous CSC contacts
Spring 1971	approx. 2000/28	Directors of corporate research
March 1971	approx. 5000/32	ASIS Members and previous CSC contacts
November 1971	approx. 60/NA	IIT trustees
November 1971	approx. 5000/10	ASIS Members and previous CSC contacts
October 1971 - February 1972	approx. 1800/48	Selected Standard Industrial Classifications with over 1000 employees in 13 midwestern states
February 1972	approx. 5000/22	ASIS Members and previous CSC contacts

10.5.2 Press Releases

Several announcements have been made to the press to publicize the Computer Search Center. In addition to newspapers and magazines that circulate to the general public, copies of the releases were sent to scientific and engineering journals in order to inform people involved with the communication of scientific information about activities of the Computer Search Center. Dates and subjects of the releases are described below.

Date	Subject
July 1970	Initiation of CSC subscriptions
November 1970	Workshop on "Computer Retrieval of Chemical and Biological Information"
March 1971	Workshop on "Computer Retrieval of Chemical and Biological Information"
Summer 1971	Advantages found by users of CSC SDI service
November 1971	Workshop on "Computer Retrieval of Scientific Information"
January 1972	Workshop on "Computer Retrieval of Scientific Information"

10.5.3 Surveys

10.5.3.1 IEEE REFLECS Survey

A questionnaire was mailed to a sample of subscribers of journals published by the Institute of Electronics and Electrical Engineers and to a sample of IEEE members in the greater Chicago area. The questionnaire and descriptive literature about the REFLECS tape were prepared in collaboration with the Information Division of IEEE.

A great deal of interest in the tape was shown by respondents. Of 89 respondents, nearly 80 percent were interested in a current awareness alerting program although a financial commitment could not be made in most cases. Respondents replied anonymously unless they were interested in follow-up information, and 73 percent elected to provide names and addresses for further information.

Although IEEE later decided not to produce the REFLECS tape, the information and insights obtained from the questionnaire were used in developing and marketing services aimed at the engineering market.

10.5.3.2 Food Technology Survey

A telephone survey of 13 major food companies was conducted in five midwest states (Illinois, Minnesota, Wisconsin, Missouri, and Michigan) to assess the degree of interest in the International Food Information Service (IFIS) data base, Food Science and Technology Abstracts. Fifteen people in 13 organizations responded. Of the 15, nine were favorable, five negative, and one undecided. Discussions are currently taking place with the Institute of Food Technologists regarding the establishment of IITRI as one of the two centers in the U.S. to handle IFIS tapes.

10.5.3.3 Market Survey

A market survey was made in 1970 to estimate the number of potential subscribers to the services of the Computer Search Center and to determine the interest in various data bases as a guide to Center expansion. The objective was to assess the potential user market in terms of size, location, experience, knowledge of data bases, preference for data bases, knowledge of computer information services, preference for information services and willingness and likelihood of paying for desired services. Because of the Center's existing services and current concentration in the chemical and biological fields, the survey concentrated primarily on the "Chemicals and Allied Products" industry. Universities, hospitals, and "Food and Kindred Products" industries were also surveyed. The survey was based upon statistical sampling.

An analysis of the distribution of chemical process plants by region and state and manufacturing employment by industry revealed that Illinois is representative not only of the East North Central Region but also the U.S. As approximately 70 percent of all industrial activity within the state of Illinois is located within the Chicago Standard Metropolitan Statistical Area, data collected within this area were considered to be representative of the state, the East North Central Region, and the United States.

Data were collected by in-depth personal and telephone interviews based on a Field Interview Guide prepared by the Center staff and Philip D. Wittlinger, Jr., of Kalish, Wittlinger and Associates, who conducted the survey. A copy of the Field Interview Guide appears in the following pages as Figures 10-34 to 10-39. A member of the Center participated in the field interviews so that the survey and subscription effort were combined to elicit information and to offer services at the same time. The survey data were used in establishing rate structures.

Although selection of organizations for interviewing had been planned on a random basis, two factors necessitated a change in the selection technique. (1) The American Petroleum Institute commenced marketing its SDI service using CA Condensates. As most all petroleum and petrochemical companies are members of the API and have financial obligations and loyalty ties with the API, it was decided not to interview them during this program because their data inputs could bias our results. (2) Twenty-six organizations with fewer than 100 employees that were contacted for the purpose of scheduling a personal interview, indicated that they had no need for an SDI service. They either did not have an R & D activity, or simply did not utilize literature search techniques within their operations.

On the basis of telephone contacts, and upon analysis of the number of employees within the Computer Search Center's client companies--all of which were organizations of over 100 employees--it was decided that organizations with fewer than 100 employees offered virtually no potential and should be excluded from further study in the survey. Thus, organizations within the petroleum/petrochemical industry and those with fewer than 100 employees were eliminated from the survey.

1. CONTACT ORIGINATED ___ TELEPHONE ____ DIRECT MAIL ___ REFERRAL ____ MEETING ___ ARTICLE

____

____

2. ORGANIZATION NAME________________________ 3. GEOGRAPHIC LOCATION______________

4. NAMES, TITLES, AND PHONE NOS. OF PERSONS INTERVIEWED

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

________________ ________________ ________________

5. PRIMARY SIC__________ 6. NO. EMPLOYEES__________ 7. SALES $______________

TYPE OF ORGANIZATION

8. ___INDUSTRIAL

9. ___ RESEARCH A.__ UNIVERSITY AFFILIATED B.__ INDEPENDENT C.__ GOVERNMENT D.___INDUSTRIAL

10.___ UNIVERSITY A. __ STATE B. __ PRIVATE

11. ORGANIZATION HAS A TECHNICAL LIBRARY? A __ YES B. __ NO

12. LIBRARY IS: A.___ CENTRAL B.___ DEPARTMENTALIZED C.___ COMBINATION A & B

13. ORGANIZATION HAS: A.____INFORMATION SCIENTIST(S) B.____LIBRARIAN(S) C.____ASS'T LIBRARIAN(S)

14. PROFESSIONAL RESEARCH STAFF IS COMPOSED OF:

A.__________CHEMISTS B.____________ C.__________ D______________

15. ORGANIZATION WAS AWARE OF THE AVAILABILITY OF COMPUTERIZED INFORMATION SYSTEM(S) YES___ NO____

Figure 10-34

FIELD INTERVIEW GUIDE - p. 1

16. **ORGANIZATION:** A. ○ CURRENTLY USING A C.I.R. SYSTEM ______________________
Name system & No. profiles

B. ○ HAVE USED A C.I.R. SYSTEM OCCASIONALLY ______________________
Name system & No. profiles

C. ○ HAVE USED A C.I.R. SYSTEM, BUT DISCONTINUED ______________________
Explain - name system, number -profiles, etc.

____

____

D. ○ HAVE INTERNAL C.I.R. SYSTEM

E. ○ HAVE NEVER USED C.I.R. SYSTEM

F. ○ CURRENTLY CONSIDERING PURCHASE OF A C.I.R. SYSTEM

G. ○ CONSIDERING DEVELOPMENT OF INTERNAL C.I.R. SYSTEM

H. ○ INTERESTED IN PERSONAL LIBRARY ON DISK

17. **LIBRARY CONTAINS** FOLLOWING ABSTRACTING JOURNALS:

(REFERENCE ATTACHED LIST)

18. RESEARCHERS WORK PRIMARILY: A. ○ IN GROUPS B. ○ INDEPENDENTLY C. ○ COMBINATION A & B

19. ORGANIZATION IS PRIMARILY INTERESTED IN A. ○ INDIVIDUAL PROFILES B. ○ GROUP PROFILES

20. INDIVIDUAL PROFILE AREAS OF INTEREST ______________________
(SPECIFY)

____

21. GROUP PROFILE AREAS OF INTEREST ______________________
(SPECIFY)

____

Figure 10-35

FIELD INTERVIEW GUIDE - p. 2

22. ORGANIZATIONS RANKING, BY THEIR UTILITY/NEED, OF DATA BASES (REFER TO MASTER LIST AND SHOW NUMBER ONLY)

A________

B________

C________

D________

E________

F________

G________

H________

I________

23. HOW MUCH TIME IS SPENT EACH WEEK IN OBTAINING TECHNICAL INFORMATION

A. LIBRARY/INFORMATION STAFF________________ B. RESEARCHERS________________

24. IN ORDER OF THEIR IMPORTANCE, RANK THE INFORMATION SOURCES RELIED UPON

A._____TECHNICAL BOOKS

B._____TECHNICAL PERIODICALS

C._____PERSONAL CONTACT

D._____PERSONAL OR COMPANY FILES

E._____HANDBOOKS; ENCYCLOPEDIAS

F._____OTHER________________ (Specify)

G._____ABSTRACTING JOURNALS

H._____TECHNICAL REPORTS

PURCHASE MECHANICS/COST JUSTIFICATION

25. REQUESTS FOR C.I.R. SERVICE FUNDING

A.____MAY BE MADE ANYTIME DURING FISCAL YEAR

B.____MUST BE INCLUDED IN BUDGET SUBMISSION

(EXPLAIN EXCEPTIONS AND/OR GENERAL COMMENTS FOR A & B________________________

____

____

26. WHAT BUDGET WOULD A C.I.R. SERVICE BE FUNDED FROM?

A. __ LIBRARY

B. __ OVERHEAD

C. ___ RESEARCH PROJECT(S)

D. ___ DEPARTMENTAL

Figure 10-36
FIELD INTERVIEW GUIDE - p. 3

27. EXPLAIN THE PROCEDURE AND IDENTIFY THE FACTORS THAT WOULD BE UTILIZED TO COST JUSTIFY THE PURCHASE OF A C.I.R. SERVICE(S)____________________

28. ORGANIZATION'S OPINION AS TO REASONABLE COST FOR SERVICE____________________

29. ________NUMBER OF PROFILES WHICH COULD BE PURCHASED BY ADJUSTING EXISTING BUDGETS.

30. ORGANIZATION'S EVALUATION OF C.I.R. SYSTEM CHARACTERISTICS

GENERAL

	ESSENTIAL	BENEFICIAL	UNIMPORTANT	NO OPINION
REGULARITY	____	____	____	____
TIMELINESS	____	____	____	____
CONSISTENCY	____	____	____	____
THOROUGHNESS	____	____	____	____
LABOR SAVING	____	____	____	____
COVERAGE	____	____	____	____
COST REDUCTIONS				
A. LABOR	____	____	____	____
B. PUBLICATIONS	____	____	____	____
C. OTHER	____	____	____	____

Figure 10-37

FIELD INTERVIEW GUIDE - p. 4

DISTINCTIVE CHARACTERISTICS OF IITRI'S CSC SYSTEM

	SIGNIFICANT	WORTHWHILE	UNIMPORTANT	NO OPINION
PROXIMITY TO CENTER	____	____	____	____
NO COST PROFILE CHANGE	____	____	____	____
LOW COST PROFILE SWITCH	____	____	____	____
MULTIPLE COPY OUTPUT	____	____	____	____
MULTILITH OUTPUT	____	____	____	____
FREE FORM BOOLEAN LOGIC	____	____	____	____
SORTING OPTIONS	____	____	____	____
RT. & LT. TRUNCATION	____	____	____	____
REDUNDANCY REMOVAL	____	____	____	____
AUTO. GEN. LIST OF				
A. PROFILE TERMS	____	____	____	____
B. TERM FREQUENCY	____	____	____	____
WEIGHTING	____	____	____	____
CONTENT & FORMAT OF OUTPUT CARDS	____	____	____	____
USER AIDS				
A. KEYLETTER IN CONTEXT LIST	____	____	____	____
B. TRUNCATION GUIDE	____	____	____	____
C. TERM FREQUENCY LIST	____	____	____	____
D. SEARCH MANUAL	____	____	____	____

Figure 10- 38

FIELD INTERVIEW GUIDE - p. 5

MATERIALS LEFT WITH ORGANIZATION

____ SEARCH MANUAL
____ BROCHURE
____ 3-PAGE INFO. SHEETS
____ OUTPUT CARDS
____ OTHER

MATERIALS PROMISED FOR FOLLOW UP

____ SEARCH MANUAL
____ BROCHURE
____ 3-PAGE INFO. SHEETS
____ OUTPUT CARDS
____ OTHER

PROFILES OFFERED FOR FREE TRIAL

No. Profiles	Data Base(s)	No. Months
____	____	____
____	____	____
____	____	____
____	____	____
____	____	____

NOTES

Figure 10- 39

FIELD INTERVIEW GUIDE - p. 6

Thirty organizations were surveyed and 70 individuals were interviewed during the course of the study. All organizations had technical libraries and in most cases they were centralized. Most organizations employed either a library staff or information scientists. Twenty-six percent of the chemical-allied products companies and 40 percent of the hospitals did not employ a special staff for literature searching and dissemination. Most organizations were aware of computer information services although only 16 percent of chemical-allied product organizations, 50 percent of the universities, and 40 percent of the hospitals were currently using a current awareness alerting (SDI) service. Occasional use of such services was reported by 16 percent of the chemical-allied product organizations and 60 percent of the hospitals. Ten percent of the former category had in-house systems and another 21 percent were considering installation of in-house systems.

Organizations that expressed little or no interest in SDI services were disinterested for one or more of the following reasons: (1) R & D efforts were in subject areas for which there is no currently-available data base; (2) R & D efforts were highly or totally applications oriented; or (3) organization compounded or blended products based upon R & D efforts of the supplier of the components of the products.

Abstracting journals were rated first by a majority of respondents in all categories. Technical serials were ranked second by a majority of respondents in all categories but the hospitals where technical books were ranked second as an information source by 40 percent of the respondents.

An evaluation of general characteristics of an SDI system was made by respondents and their responses are summarized in Table 10-25. Labor saving, coverage, and thoroughness were considered to be essential characteristics of a system by a plurality of respondents.

Respondents also evaluated specific characteristics of

Characteristic	Chemical-Allied Products				Food-Kindred Products				Universities				Hospitals			
	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION
Regularity	32	42	5	21	33	67	-	-	50	25	25	-	20	40	-	40
Timeliness	37	21	21	21	-	100	-	-	50	25	25	-	20	40	-	40
Consistency	32	37	5	26	67	33	-	-	50	50	-	-	40	20	-	40
Thoroughness	64	10	5	21	67	-	33	-	75	25	-	-	60	-	-	40
Labor Saving	69	5	5	21	67	33	-	-	75	25	-	-	60	-	-	40
Coverage	43	26	5	26	100	-	-	-	75	25	-	-	60	-	-	40
Cost reduction - labor	26	32	10	32	33	33	34	-	25	50	-	25	20	20	20	40
Cost reduction - publications	16	21	26	37	-	33	67	-	-	25	50	25	-	-	60	40

* Totals exceed 100% because of multiple responses.

Table 10-25

EVALUATION OF SDI SYSTEM CHARACTERISTICS*

(PERCENT OF RESPONDENTS)

IITRI's Computer Search System as these were described by interviewers. In considering Table 10-26, it should be borne in mind that the tabulated evaluations are based upon anticipation and not working experience with the system. Significant characteristics included proximity to the center, no cost profile change, low cost profile switch, free form Boolean logic, truncation, and content and format of output cards. Multiple copy output and multilith output were rated unimportant by a majority of respondents.

Characteristic	Chemical-Allied Products				Food-Kindred Products				Universities				Hospitals			
	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION	ESSENTIAL	BENEFICIAL	NOT IMPORTANT	NO OPINION
Proximity to Center	42	5	32	21	67	33	-	-	25	25	50	-	80	-	-	20
No cost profile change	58	10	-	32	100	-	-	-	50	50	-	-	60	-	-	40
Low cost profile switch	53	10	5	32	100	-	-	-	100	-	-	-	20	40	-	40
Multiple copy output	5	16	47	32	33	-	67	-	-	-	100	-	-	-	60	40
Multilith output	5	10	53	32	33	-	67	-	-	-	100	-	-	-	60	40
Free form Boolean logic	63	5	5	27	34	33	-	33	100	-	-	-	40	20	-	40
Sort options	26	21	21	32	-	100	-	-	25	50	-	25	20	40	-	40
Truncation-left and right	47	21	5	27	34	33	-	33	75	25	-	-	40	20	-	40
Redundancy removal	10	32	5	53	-	33	-	67	25	-	25	50	-	40	-	60
Weighting	42	16	5	37	34	33	33	-	50	50	-	-	20	40	-	40
Content-format of output cards	58	21	-	21	34	33	33	-	75	25	-	-	40	20	-	40

* Totals exceed 100% because of multiple responses.

Table 10-26

EVALUATION OF CHARACTERISTICS OF IITRI'S RETRIEVAL SYSTEM*

(PERCENT OF RESPONDENTS)

10.5.4 Pricing

Our current subscription fees are shown on the cost sheets for CA, BA and EI (Figures 10-40, 10-41, and 10-42, respectively). Initially, we had based our charges on a profile rather than on the present system of input and output units. However, since there were no restrictions on the size of an individual profile, there was an imbalance between our cost and the fees we charged. Profiles of two terms cost much less to run than profiles of two hundred terms, yet the subscription fees were the same. Compounding the problem was the fact that some economy-minded users took advantage of the free-form Boolean logic capability to ask several questions in one huge profile. To combine three questions, they merely had to put each separate question's logic expression within parentheses and "0R" the three sets together. Weights could be used to segment the output into three sets. An evaluation made after several months of charging under the profile-based system indicated that 10% of the users, paying 10% of the fees, accounted for more than 40% of our costs.

After the first year of operation, we changed our fee structure to one based on units of input (search terms) and units of output (citations printed). This system more closely reflects our actual costs. No limitations to profile size are necessary and, if desired by the user, several questions can be combined in one profile. However, the cost will reflect the profile's size and number of citations retrieved. Since our statistics showed that over 75% of the profiles could be coded in 25 terms or less and would retrieve 50 or fewer citations per issue searched, we set our basic input unit at 25 terms and our basic output unit at 50 citations retrieved per issue searched. Supplemental units are based on each unit of 1-10 search terms for input, and 1-50 citations retrieved per issue searched for output. Both input and output units are averaged over the subscription period to even out minor fluctuations. We also give discounts for several profiles mailed to one address, reflecting our decreased handling costs for those cases.

This subscription fee system is more equitable. Some users receive more service than others since they request changes more often, but we do not plan to charge for revisions. We think that such a charge might stifle legitimate reasons for change and denigrate profile performance. We have an accounting program to keep track of search terms used and output generated for each profile, so the system is not cumbersome to operate. Although the rates may change as data base sizes increase and costs go up, we will probably retain this basic structure.

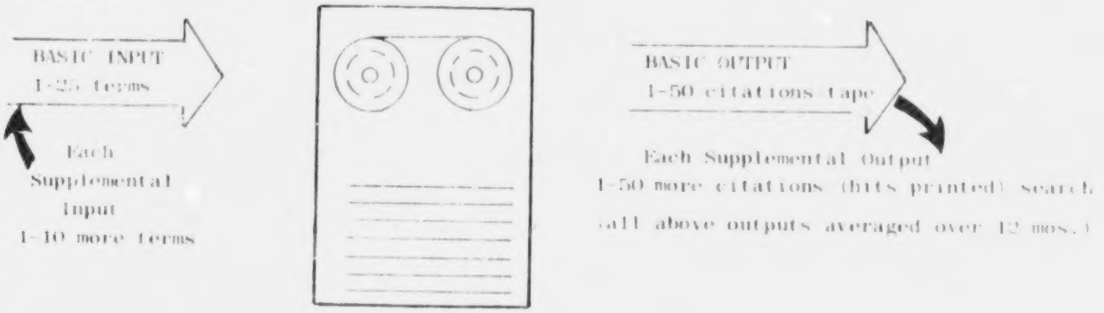
COMPUTER SEARCH CENTER

at IIT Research Institute

CHEMICAL ABSTRACTS CONDENSATES (CAC)

Chemical Abstracts Service issues a CAC tape weekly. Each CAC tape corresponds to the weekly printed issue of Chemical Abstracts. Twenty-six weeks (issues) of CAC comprise one volume; two volumes are published yearly. Odd numbered tapes cover sections 1-34 (organic); even numbered tapes cover sections 35-80 (inorganic). CAC includes citations for each entry in CA (about 300,000 annually) which covers chemical literature throughout the world.

SUBSCRIPTION STRUCTURE


BASIC INPUT
1-25 terms
Each
Supplemental
Input
1-10 more terms
BASIC OUTPUT
1-50 citations tape
Each Supplemental Output
1-50 more citations (hits printed) search
(all above outputs averaged over 12 mos.)

ANNUAL SUBSCRIPTION RATES

CATEGORY	NUMBER ISSUES	BASIC UNIT COMBINATION	EACH SUPPLEMENTAL INPUT	EACH SUPPLEMENTAL OUTPUT
CA-1	26 (either even or odd)	$165	$ 60	$ 60
CA-2	52 (both even and odd)	$250	$100	$100

GROUP DISCOUNTS

Ten or more users within one organization (one mailing address) may subscribe at the reduced rates of $145 and $220 for CA-1 and CA-2, respectively. These rates are available immediately when ten or more users enter subscriptions within a 30 day period. If ten or more users enter subscriptions over a period longer than 30 days, their renewals will be at the discounted rate.

HOW TO SUBSCRIBE

All subscriptions should be submitted on an organization's purchase order with full prepayment.

Make checks payable to IIT RESEARCH INSTITUTE - CSC.

Mail to:
Martha E. Williams
Manager
Computer Search Center
10 West 35th Street
Chicago, Illinois 60616

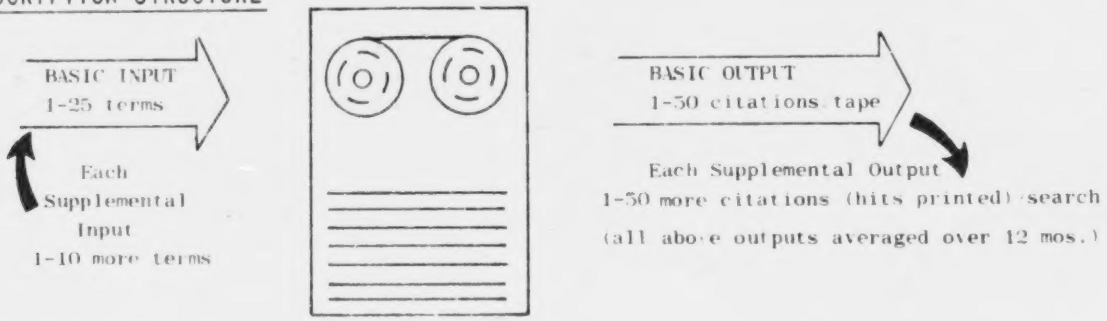
Figure 10-40
CA PRICE SHEET

COMPUTER SEARCH CENTER
at IIT Research Institute

BIOLOGICAL ABSTRACTS PREVIEWS (BA Previews)

BA (issued biweekly) covers biological journals throughout the world and provides 140,000 citations annually. BioRI (issued monthly) provides 100,000 citations annually and covers other biological publications such as symposia proceedings, government reports and conference papers.

SUBSCRIPTION STRUCTURE


BASIC INPUT
1-25 terms
Each Supplemental Input
1-10 more terms
BASIC OUTPUT
1-50 citations/tape
Each Supplemental Output
1-50 more citations (hits printed)/search
(all above outputs averaged over 12 mos.)

ANNUAL SUBSCRIPTION RATES

CATEGORY	NUMBER ISSUES	BASIC UNIT COMBINATION	EACH SUPPLEMENTAL INPUT	EACH SUPPLEMENTAL OUTPUT
BA-1	12 BioRI	$130	$ 45	$ 45
BA-2	24 BA	$200	$ 75	$ 75
BA-3	36 Both	$250	$100	$100

GROUP DISCOUNTS

Ten or more users within one organization (one mailing address) may subscribe at the reduced rates of $120, $170, and $220 for BA-1, BA-2, and BA-3, respectively. These rates are available immediately when ten or more users enter subscriptions within a 30 day period. If ten or more users enter subscriptions over a period longer than 30 days, their renewals will be at the discounted rate.

HOW TO SUBSCRIBE

All subscriptions should be submitted on an organization's purchase order with full prepayment.

Make checks payable to IIT RESEARCH INSTITUTE - CSC.

Mail to: Martha E. Williams
Manager
Computer Search Center
10 West 35th Street
Chicago, Illinois 60616

Figure 10-41
BA PRICE SHEET

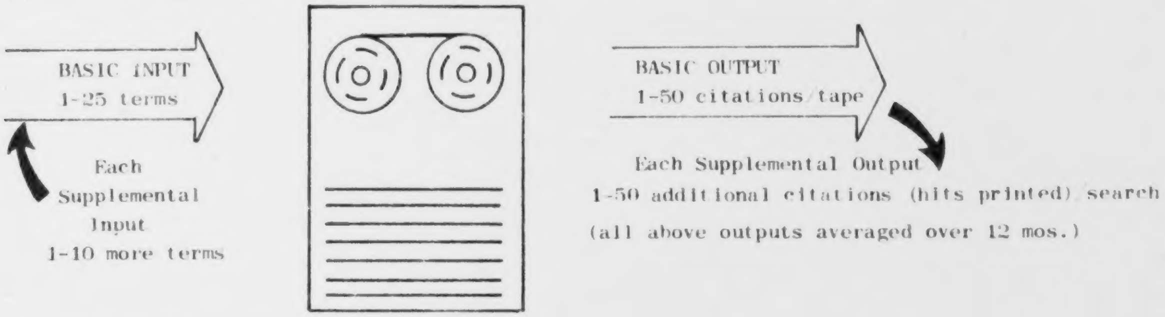
COMPUTER SEARCH CENTER

at IIT Research Institute

COMPuterized ENgineering inDEX (COMPENDEX)

Engineering Index publishes monthly the COMPENDEX tape, a compilation of key engineering journals throughout the world. Over 3500 journals, conference proceedings, and other publications are covered, providing over 84,000 citations annually.

SUBSCRIPTION STRUCTURE


BASIC INPUT
1-25 terms
Each
Supplemental
Input
1-10 more terms
BASIC OUTPUT
1-50 citations/tape
Each Supplemental Output
1-50 additional citations (hits printed) search
(all above outputs averaged over 12 mos.)

ANNUAL SUBSCRIPTION RATES

CATEGORY	NUMBER ISSUES	BASIC UNIT COMBINATION	EACH SUPPLEMENTAL INPUT	EACH SUPPLEMENTAL OUTPUT
EI-1	12	$200	$75	$25

GROUP DISCOUNTS

Ten or more users within one organization (one mailing address) may subscribe at the reduced rate of $175 for EI-1. This rate is available immediately when ten or more users enter subscriptions within a 30 day period. If ten or more users enter subscriptions over a period longer than 30 days, their renewals will be at the discounted rate.

HOW TO SUBSCRIBE

All subscriptions should be submitted on an organization's purchase order with full prepayment.

Make checks payable to IIT RESEARCH INSTITUTE - CSC.

Mail to: Martha E. Williams
Manager
Computer Search Center
10 West 35th Street
Chicago, Illinois 60616

Figure 10-42
EI PRICE SHEET

10.5.5 Brochures

Of the many types of publicity used by the Computer Search Center, workshop and CSC brochures have probably received the widest circulation. Over 5,000 brochures announcing the latest workshop on Computer Retrieval of Scientific Information were sent to people who had had previous contact with the CSC or who were known to be interested in information science. The CSC brochure is used for all general publicity mailings, since it lists CSC services and gives examples of typical output. The workshop brochure is shown in Figures 9-1 and 9-2, and the CSC brochure is shown in Figures 10-43 and 10-44.

10.5.6 Contacts

Design, implementation, and development of the Computer Search Center have resulted in a great many contacts with information scientists from other organizations, potential users, etc. Over the past four years, 1175 individuals in 719 distinct organizations have been in contact with Computer Search Center personnel. These figures represent contacts made in person, via telephone calls or via individual correspondence. Individuals contacted as a result of a direct mailing are not included in the above numbers unless they responded by requesting further information.

SERVICES

Selective dissemination of information (current awareness alerting)

Retrospective searches

Workshops

Seminars

Personal libraries stored on computer

QUERIES

Personal profiles

Group profiles

SEARCH PARAMETERS

Wide range of access terms

Left and right truncation

Free-form Boolean logic

Weights

CUSTOM OUTPUT

Cards or paper listing

Sort by: reference number
author
weight

SECURITY is provided for proprietary information.

COMPUTER SEARCH CENTER

For more information
contact:

Computer Search Center
IIT Research Institute
10 West 35 Street
Chicago, Ill. 60616

Phone: 312/225-9630
ext. 4918

COMPUTER SEARCH CENTER

A one stop INFORMATION CENTER to answer the needs of:

INDUSTRY
RESEARCH ORGANIZATIONS
EDUCATIONAL INSTITUTIONS

Programs have been designed to search a wide variety of source tapes which are converted to a standard format for searching on IBM 360 series computers.

DATA BASES currently available or planned for the future include:

CHEMICAL ABSTRACTS

CONDENSATES
CBAC (chemical-biological activities)
POST (polymer science and technology)
SSS (substructure search system)

BIOLOGICAL ABSTRACTS

BA PREVIEWS

ENGINEERING INDEX

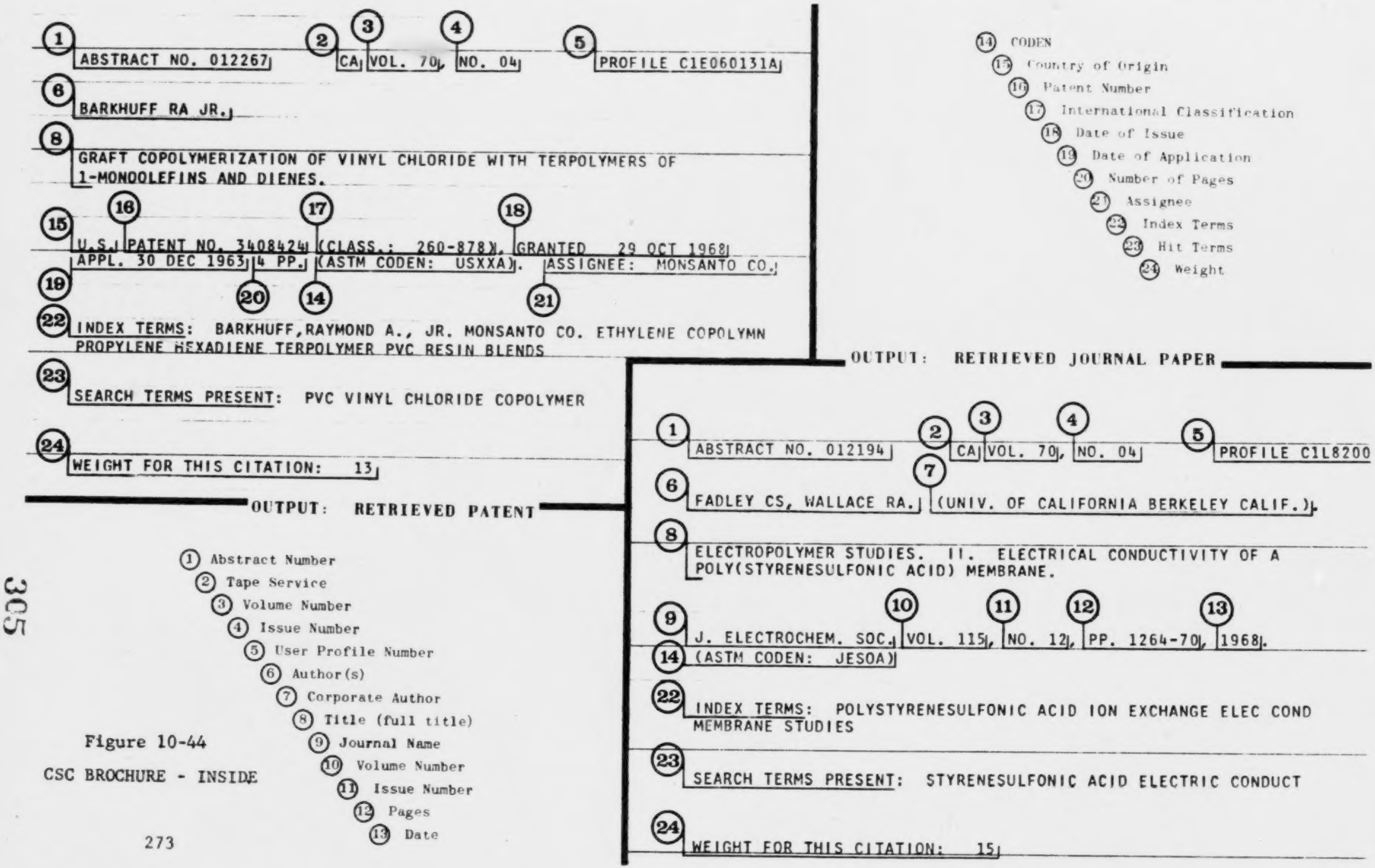
COMPENDEX

INSTITUTE FOR SCIENTIFIC INFORMATION

ASCA (automatic subject citation alerting, including source and citation tapes)

Figure 10-43

CSC BROCHURE - OUTSIDE



```
ABSTRACT NO. 012267   CA VOL. 70, NO. 04        PROFILE C1E060131A
BARKHUFF RA JR.
GRAFT COPOLYMERIZATION OF VINYL CHLORIDE WITH TERPOLYMERS OF
1-MONOOLEFINS AND DIENES.
U.S. PATENT NO. 3408424 (CLASS.:  260-878), GRANTED   29 OCT 1968
APPL. 30 DEC 1963 4 PP. (ASTM CODEN:  USXXA).  ASSIGNEE:  MONSANTO CO.
INDEX TERMS:  BARKHUFF,RAYMOND A., JR. MONSANTO CO. ETHYLENE COPOLYMN
PROPYLENE HEXADIENE TERPOLYMER PVC RESIN BLENDS
SEARCH TERMS PRESENT:  PVC VINYL CHLORIDE COPOLYMER
WEIGHT FOR THIS CITATION:   13
```

OUTPUT: RETRIEVED PATENT

```
ABSTRACT NO. 012194   CA VOL. 70, NO. 04        PROFILE C1L8200
FADLEY CS, WALLACE RA. (UNIV. OF CALIFORNIA BERKELEY CALIF.).
ELECTROPOLYMER STUDIES.  II.  ELECTRICAL CONDUCTIVITY OF A
POLY(STYRENESULFONIC ACID) MEMBRANE.
J. ELECTROCHEM. SOC. VOL. 115, NO. 12, PP. 1264-70, 1968.
(ASTM CODEN:  JESOA)
INDEX TERMS:  POLYSTYRENESULFONIC ACID ION EXCHANGE ELEC COND
MEMBRANE STUDIES
SEARCH TERMS PRESENT:  STYRENESULFONIC ACID ELECTRIC CONDUCT
WEIGHT FOR THIS CITATION:   15
```

OUTPUT: RETRIEVED JOURNAL PAPER

① Abstract Number
② Tape Service
③ Volume Number
④ Issue Number
⑤ User Profile Number
⑥ Author(s)
⑦ Corporate Author
⑧ Title (full title)
⑨ Journal Name
⑩ Volume Number
⑪ Issue Number
⑫ Pages
⑬ Date

⑭ CODEN
⑮ Country of Origin
⑯ Patent Number
⑰ International Classification
⑱ Date of Issue
⑲ Date of Application
⑳ Number of Pages
㉑ Assignee
㉒ Index Terms
㉓ Hit Terms
㉔ Weight

Figure 10-44

CSC BROCHURE - INSIDE

10.6 Contacts and Cooperative Arrangements

ASIDIC, the Association of Scientific Information Dissemination Centers, was begun September 18-19, 1968. At that time, representatives of various centers providing services from machine-readable data bases developed by Chemical Abstracts Services met at CAS to discuss their mutual goals and problems. Members of IITRI's Computer Search Center were active at this formative meeting. A series of workshops followed. They were held at IITRI (November 13, 1968), the University of Georgia (August 26-28, 1968 and February 27-28, 1969), and the University of Pittsburgh (June 17-18, 1969) and dealt with programming, profile development and inter-center relationships. By mid-1969, the group had grown both in size and interests, as many industrial, university, and not-for-profit organizations were involved in processing a variety of data bases.

On October 22-23, 1969, ASIDIC offically came into being with the election of officers and development of a charter. Eugene Schwartz of IITRI served as the first president of ASIDIC. A pattern of two annual meetings developed. One, open to all, is devoted to annual business and items of general interest. The second retains the flavor of the earlier workshops and provides an opportunity for small group round-table discussions of common problems. The official purposes of ASIDIC are:

- to promote applied technology of information storage and retrieval as related to large data bases containing bibliographic, textual and fact information
- to share experience and information through meetings, seminars and workshops
- to recommend standards for data elements, formats and codes
- to promote research & development for more efficient use of varied data bases.

Full membership is reserved for centers providing services to over 100 users from two or more data bases (not internal).

IITRI has maintained a continued interest in and service to ASIDIC. Martha Williams is the current Vice President, a member of the Committee on Center-Supplier Relations, and chairman of the Cooperative Data Management Committee, which recently compiled an extensive survey of centers and services. Peter Schipma has been an active member on the Standards Committee since its inception.

Over 20 data base suppliers and a similar number of centers, universities, industrial organizations, and government agencies have been contacted concerning possible data base use or informal networking. These discussions are continuing at the present time. Foreign countries with which contacts have been made include:

Argentina
Austrailia
Austria
Belgium
Brazil
Canada
Ceylon
Chile
Czechoslovakia
Denmark
England
Finland
France
Germany
Hungary
India
Ireland
Israel
Italy
Japan
Korea
Mexico
Netherlands
Spain
Sweden
Thailand
Union of South Africa

11. RESEARCH STATISTICS, COMPUTATIONAL LINGUISTICS AND ANALYTICAL STUDIES

In order to provide good service to users and to gain insights that may lead toward future developments within or related to CSC, we maintain statistics on and conduct research related to various aspects of users, data bases, systems, and personnel. Statistics and records are maintained and research is conducted in an effort to:

improve profiles
monitor user response
monitor data bases
improve methods of using data bases
suggest improvements for data bases
observe trends
devise cost accounting procedures
monitor program efficiencies
improve search strategies
obtain data for future planning
improve system
monitor and project personnel needs
generate data for further study.

In addition to the data base statistics provided in Section 6 and production statistics in Section 10, we maintain a variety of statistics on system features, profile terms, profiles, and hits (output).

11.1 System Features

The CSC system includes certain design features which were employed following a study of the desirable and desired features of systems we analyzed during the design phase of C6156. We have since analyzed CSC profiles to determine the extent of use of the design features: linking, truncation, variable term types, free form Boolean logic, and weights.

11.1.1 Linking of Terms (See Section 4.2.3)

Links or groups are employed extensively by users of the IITRI system. For example, in a typical run against an odd numbered issue of CA Condensates, 94% of all the terms used in all profiles were included in links. Only 6% of the terms were referred to individually in the profile logic. While 6% of the total number of terms in the run were not in links, only 5.6%

of the profiles (user questions) used no links at all. The majority of the profiles, 73.6%, used one to four links, 20.8% used five to ten links, and none used more than ten.

The number of terms in a single link has varied from one to 120. However, a more normal range is the range of one to 38 observed in the run under discussion. The average number of terms per link is eight.

11.1.2 Truncation and Various Data Types

Truncation can be used with any kind of data element or term type in a given data base. An analysis of the use of the various truncation modes (none, left, right, and both left and right simultaneously) versus term type, indicates that, when searching an issue of CA, the search terms that users truncate are text terms (index and title terms), author terms, CODEN, CA section numbers, and corporate authors. As one might assume, subject terms or text terms comprise the majority of the terms in profiles, followed by CA Section number, CODEN, author and corporate author. In fact, by term type, 93.2% of the terms were subject or text terms, 2.0% were authors, 2.3% were CA section numbers, 2.1% were CODEN, and 0.4% were corporate authors.

Table 11-1 gives the numbers of terms and various term types vs. truncation modes used in a particular run.

Naturally, right truncation is the most commonly used truncation mode. As can be seen in Table 11-1, of the text terms, 54.8% are right truncated, 26.3% are not truncated at all, 16.3% have simultaneous left and right truncation, and 2.6% are left truncated. Note that the individual left and right truncation modes do not include the instances of both left and right truncation, hence if one wanted to know all instances of left truncation, and not merely left and only left, he could add the numbers from the "both" line to the numbers for left truncation (and similarly to the numbers for right truncation.) Thus, using the numbers in Table 11-1 for text terms, all instances of left truncation would be 18.9% (2.6+16.3) and all instances of right truncation would be 71.1%(54.8 + 16.3). CA section

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

	Term Type				
Truncation Mode	Coden	CA Section	Text	Author	Corporate Author
None	102 (65.0%)	28 (16.2%)	1811 (26.3%)	48 (32.0%)	22 (84.6%)
Left	0	28 (16.2%)	179 (2.6%)	0	0
Right	1 (0.6%)	20 (11.6%)	3768 (54.8%)	102 (68.0%)	4 (15.4%)
Both	54 (34.4%)	97 (56.0%)	1119 (16.3%)	0	0

Table 11-1

NUMBER OF TERMS OF VARIOUS TERM TYPES
VS. TRUNCATION MODE USED

numbers and corporate authors are either right truncated or not truncated at all. Left truncation would be of no meaningful use. Right truncation on a CA section number would allow a user to pick up 10 sections in biochemistry with the single truncated term CA01*. CA01* will cover sections 10 through 19.

When truncation is used with author names it is usually right truncation and is helpful in picking up names that are spelled differently in a foreign language and transliterated in several ways. Left truncation on an author name will retrieve variant representations of names such as O'Hara where the spacing between the "O" and the "H" might vary and the punctuation might be included in some cases and not others.

In the case of CODEN, truncation is little used but valuable when needed. There is no need to truncate the CODEN for a specific journal, in fact to do so would provide false retrieval. In the case of conferences and proceedings, which are designated by a one or two in the first position of the CODEN, right truncation can be used. Simultaneous left and right truncation on patent CODEN is used. The third and fourth positions in the CODEN for patents are designated XX, and one can use the truncated search term *XX* to retrieve all patent references.

Table 11- 2 shows the number of profiles, in a run, containing various term types with the truncation modes used. Table 11- 3 shows the percent of profiles containing the various term types versus the truncation mode used.

Truncation has been employed by all of the participants in the CSC SDI program. Considering all the profile terms in several runs:

No truncation was used for 46% of the terms
Left truncation was used for 5% of the terms
Right truncation was used for 36% of the terms
Both truncation was used for 13% of the terms

---

* - Denotes truncation

These statistics, initially generated on a computer-manual basis, are now completely machine generated. (See Table 11-2).

```
RESULTS OF TERM PROCESSING

 3500 TERMS
 2630 UNIQUE TERMS
  450 LCBS USED (OUT OF  2003 )

MEAN FREQ. OF TERM LCBS IS      19693.317
S.D. OF FREQ. OF TERM LCBS IS   8316.434

MEAN FREQ. OF ALL LCBS IS        8410.566

MEAN GROUP SIZE IS        5.843
S.D. OF GROUP SIZE IS     5.549

ALL MEANS AND S.D.S BASED ON UNIQUE TERM COUNT



CROSS-TABULATION OF TERM TYPE VS MODE OF TRUNCATION

TYPE -      1     2     3     4     5     6     7     8     9    10    11    12    13    14    15 TOTAL

MODE 0     43   873    23     0     0     0     0    11     0     1     0     0     0     0     0   951

MODE 1      0    73     0     0     0     0     0     0     0     0     0     0     0     0     0    73

MODE 2      0  1778    45     0     0     0     0     3     0     0     0     0     0     0     0  1826

MODE 3     29   621     0     0     0     0     0     0     0     0     0     0     0     0     0   650

TOTAL      72  3345    68     0     0     0     0    14     0     1     0     0     0     0     0  3500


CROSS-TABULATION BASED ON UN-AGGREGATED TERMS.
```

Table 11-2

STATISTICAL OUTPUT FROM INPUTR
(profiles for CA77:01)

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

	Term Type				
Truncation Mode	Coden	CA Section	Text	Author	Corporate Author
None	2.23	6.34	84.32	4.47	1.86
Left	0	4.47	27.98	0	0
Right	.37	2.23	91.41	5.22	1.49
Both	2.23	13.05	76.11	0	0

Table 11-3

PERCENT OF PROFILES CONTAINING VARIOUS TERM TYPES
VS. TRUNCATION MODE USED

11.1.3 Free Form Boolean Logic

During analysis of profiles run against CA Volume 76, issues 25 and 26, we determined that 86.9% of the profiles used AND logic, 77.6% used OR logic, and 32.4% used NOT logic. Table 11-4 indicates the number of times each logic operator was used within a profile. For example, 35 profiles or 13.1% of the profiles did not use AND logic; 67 profiles or 25% used the AND operator only once; and four profiles or 1.5% of the profiles used AND ten or more times. The frequency of use of OR logic is similar to that of AND. NOT logic, while used in a larger percentage of profiles than one might suspect, is not used very frequently within a single profile. It is used in 32.5% of all profiles--once in 27.6% of the profiles, twice in 3.2% and three times in 1.1% of the profiles. The NOT operator is not used more than four times in any profile.

The CSC search system allows any number of parenthetic logic statements in a profile and they can be nested to any degree. Table 11- 5 indicates the number of sets of parentheses found in the same group of profiles. Sixty-five profiles or 24.3% used no parentheses, and 75.7% did use parentheses. Thirty-nine profiles or 14.2% used one set of parentheses, 50 profiles or 18.7% used two sets, etc. The purpose of this analysis is to indicate the fact that where permitted to use free logic the user does make use of that feature. The number of sets of parentheses is some indication of the degree of complexity and length of the search question. The actual use of nested logic is given in table 11-6.

11.1.4 Weighting

Weights were used in 24.14% of all profiles run against CA Volume 76. (24.76% for the even numbered issues and 23.52% for the odd numbered issues). This is an increase over the 11.6% use experienced in Volume 71. The reason for the increase is most likely due to our change in our basis for pricing.

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

Number of Times Logic Operator Used in a Profile	AND		OR		NOT	
	Number of Profiles	Percent of Profiles	Number of Profiles	Percent of Profiles	Number of Profiles	Percent of Profiles
0	35	13.1	60	22.4	181	67.5
1	67	25.0	59	22.0	74	27.6
2	57	21.3	46	17.2	9	3.4
3	39	14.6	15	5.6	3	1.1
4	26	9.7	23	8.6	1	.4
5	14	5.2	21	7.8	0	0
6	14	5.2	14	5.2	0	0
7	7	2.6	9	3.4	0	0
8	3	1.1	3	1.1	0	0
9	2	.7	1	.4	0	0
10 (or more)	4	1.5	17	6.3	0	0
Total	268	100	268	100	268	100
Total Number of Logic Operator Appearances	699		794		105	
Percent Use of Logic Operators Using All Operators in the Run	43.7		49.7		6.6	

Table 11-4

NUMBER AND PERCENT OF PROFILES USING AND, OR, and NOT LOGIC VS. NUMBER OF TIMES EACH OPERATOR WAS USED IN A PROFILE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

Number of Sets of Parentheses	Number of Profiles	% Profiles
0	65	24.3
1	39	14.6
2	50	18.7
3	29	10.8
4	24	9.0
5	18	6.7
6	10	3.7
7	7	2.6
8	3	1.1
9	9	3.3
10 (or more)	14	5.2
Total	268	100.0

Table 11- 5

USE OF PARENTHETIC LOGIC IN PROFILES

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

Highest Degree of Nesting of Parentheses	Number of Profiles	% Profiles
0	65	24.3
1	89	33.2
2	70	26.1
3	29	10.8
4	13	4.9
5	2	.7
Total	268	100.0

Table 11- 6

USE OF NESTED LOGIC IN PROFILES

Initially there was no limit to the number of terms a user could put in a profile. Later, when we found 10% of our users were costing 40% of the machine time, we decided to assign a term limit of 25. This encouraged users to try to use all of their 25 terms, hence a user with a one term profile would combine his with one or two other users from the same company. Because of the flexibility of the logic system they could specify three profiles as one and separate the questions with OR logic operators. Faced with the problem of combined output they would then assign zero weight to one question and two distinct weights (high and low weights) to the other questions. The net result was that the zero weighted profile's output would be printed first, the low weighted one's second, and the high weighted one's last.

11.2 Terms--Profiles

A retrieval system that involves natural language terms is bound to be term oriented, i.e., the crux of the system involves matching the intent of a user's question with the intent of a titled-indexed reference, and the match takes place through terms--either terms *per se* or terms that have been coded, truncated, classified, etc. The terms of the profile and the terms on the data base are of great importance. The profile terms are designated by the CSC profiler and/or the user, and data base terms by the supplier.

After checking the user aids in order to exercise what control we can on profile terms (term frequencies and term fraction occurrences) we prepare complete profiles incorporating appropriately truncated terms and logic, etc.

Aggregation is the preparation of one sorted list containing one occurrence only of each term from the total batch of all profile terms in a run. The larger the profile term list the greater the benefits of aggregation are, and conversely, if a term list is reduced or split into two batches for separate runs the benefits are diminished. A term that appears in several profiles appears only once in the aggregated word list together with information concerning the profiles in which the term appears. The programming aspects of aggregation have been discussed in Section 5 under the INPUTR program. Aggregation serves several purposes. It effects a savings in search time required--if a term is used in multiple profiles it need only be searched once. An alphabetical profile term list is printed out for all terms used in all profiles in a given run. This shows spelling errors in profile input that should not but occasionally do occur. It also shows variation in truncation which may be either intentional or wasteful. One cannot automatically determine where to truncate on a term, as the content of two or more profiles using common term fractions may differ, resulting either in loss of relevant information or in an overabundance of false hits. The aggregation feature was included in the initial program design in

1968 and has proved to have economic benefit. In our first production run we had only 800 profile terms before aggregation and these were reduced by 15.7% to 674 terms actually submitted for searching. When we reached 3758 profile terms, we achieved a reduction of 29.5% to 2650 terms.

The aggregation ratio is dependent on the number and character of profiles in a run. Homogeneity of profiles increases the likelihood of identical terms being used in more than one profile, and in a large number of profiles the number of occurrences of specific terms is likely to be higher. Aggregation is affected by use of Standard truncations. (See Section 7.4). Term aggregation for profiles run against issues of CA, BA, and EI are shown in Figures 11-1 through 11-9. These numbers expressed in terms of an aggregation reduction ratio are presented in Figure 11-10 through 11-18. The average number of terms per profile vs. issues of CA, BA, and EI are given in Figures 11-19 through 11-27. The average number decreased once the free pilot runs terminated and the subscription fees were introduced. The average reached in Volume 72 was 34. We announced our prices and the averages started to decrease. The current average is 24.

Cost per term vs. issue and cost per term per citation vs. issue are given for CA, BA, and EI in Figures 11-28 through 11-45. The cost per profile for each of the issues searched is given in Figures 11-46 through 11-54. The cost/profile for searches of CA have steadily decreased from approximately $11.00/issue to $1.75/issue for Volume 76. This decrease is due to continued efforts to increase the efficiency of the software.

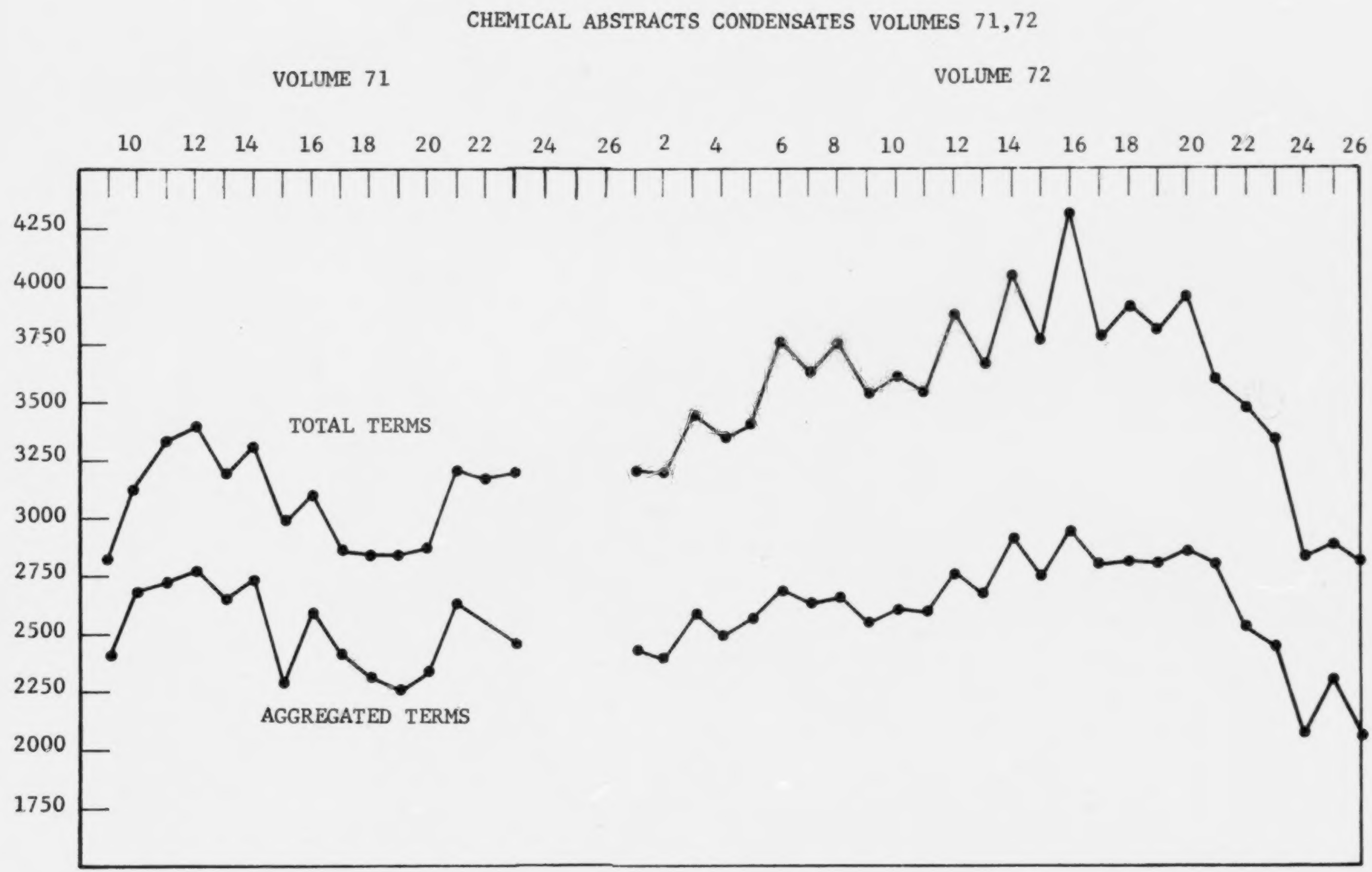


Figure 11-1

TERM AGGREGATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

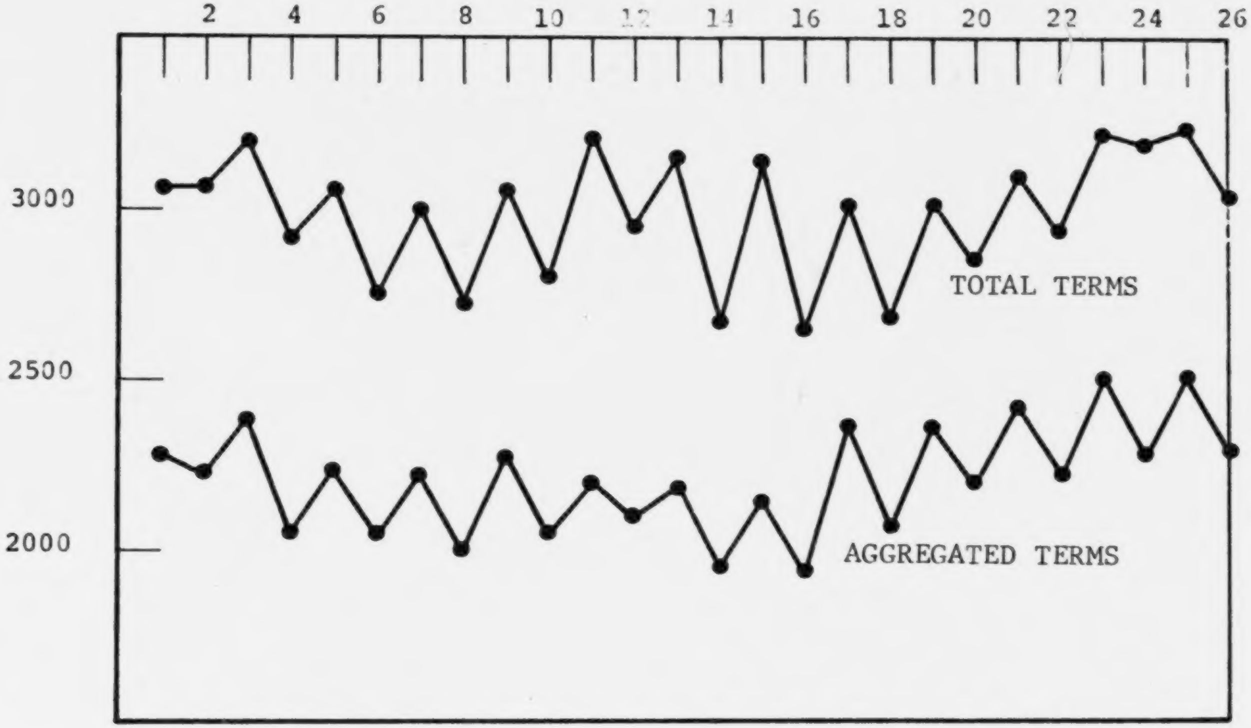


Figure 11-2

TERM AGGREGATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

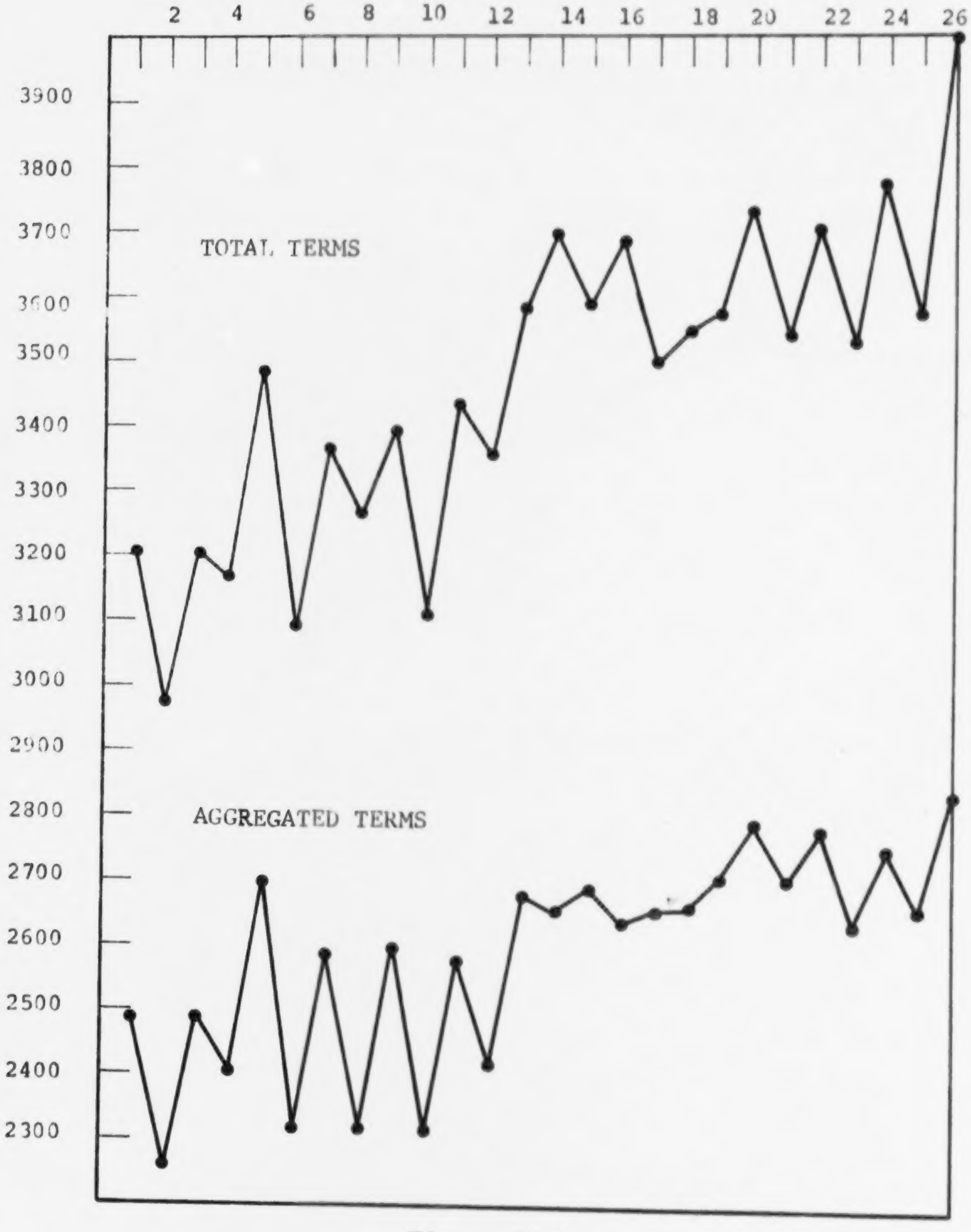


Figure 11-3

TERM AGGREGATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

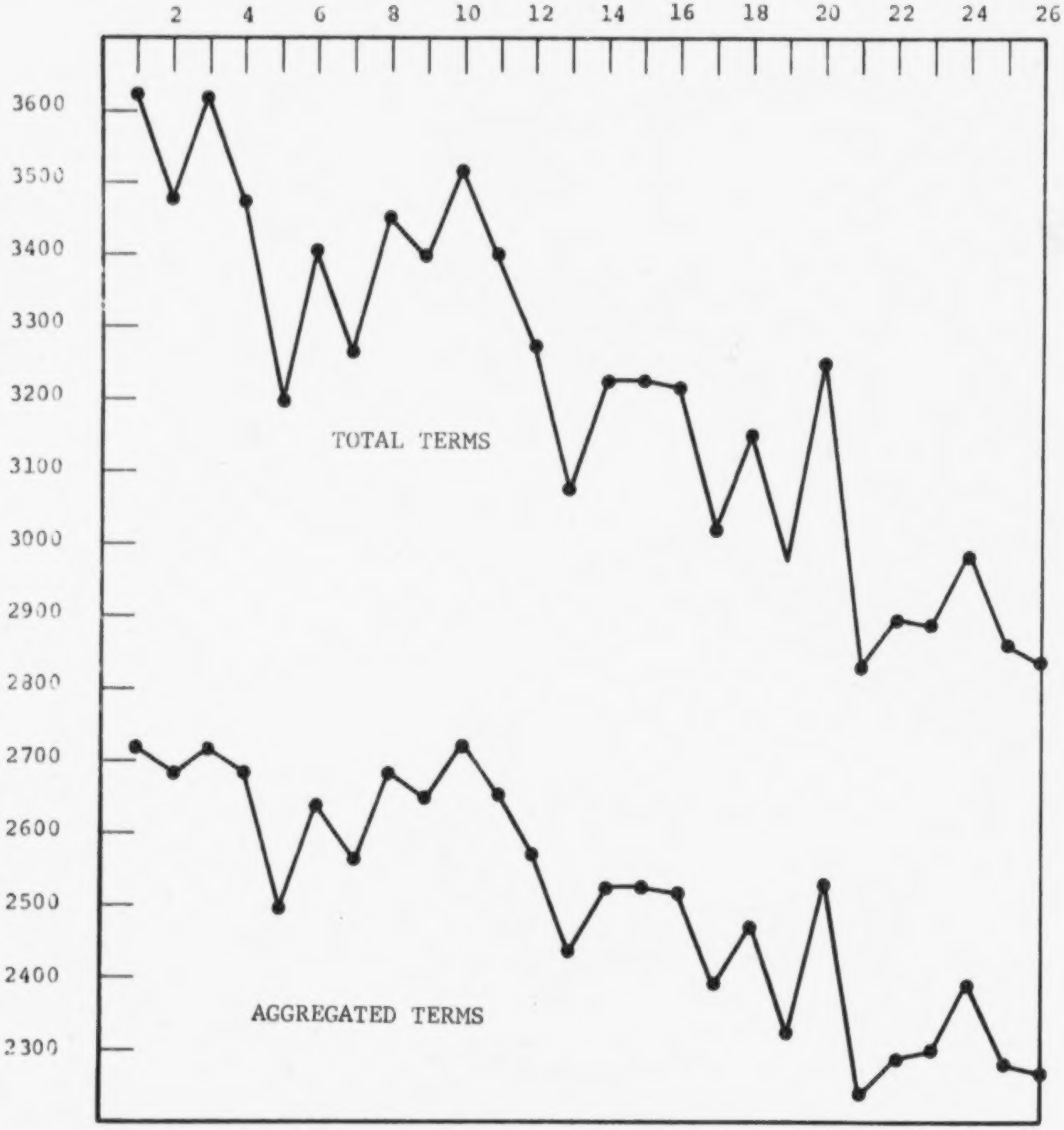


Figure 11-4

TERM AGGREGATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

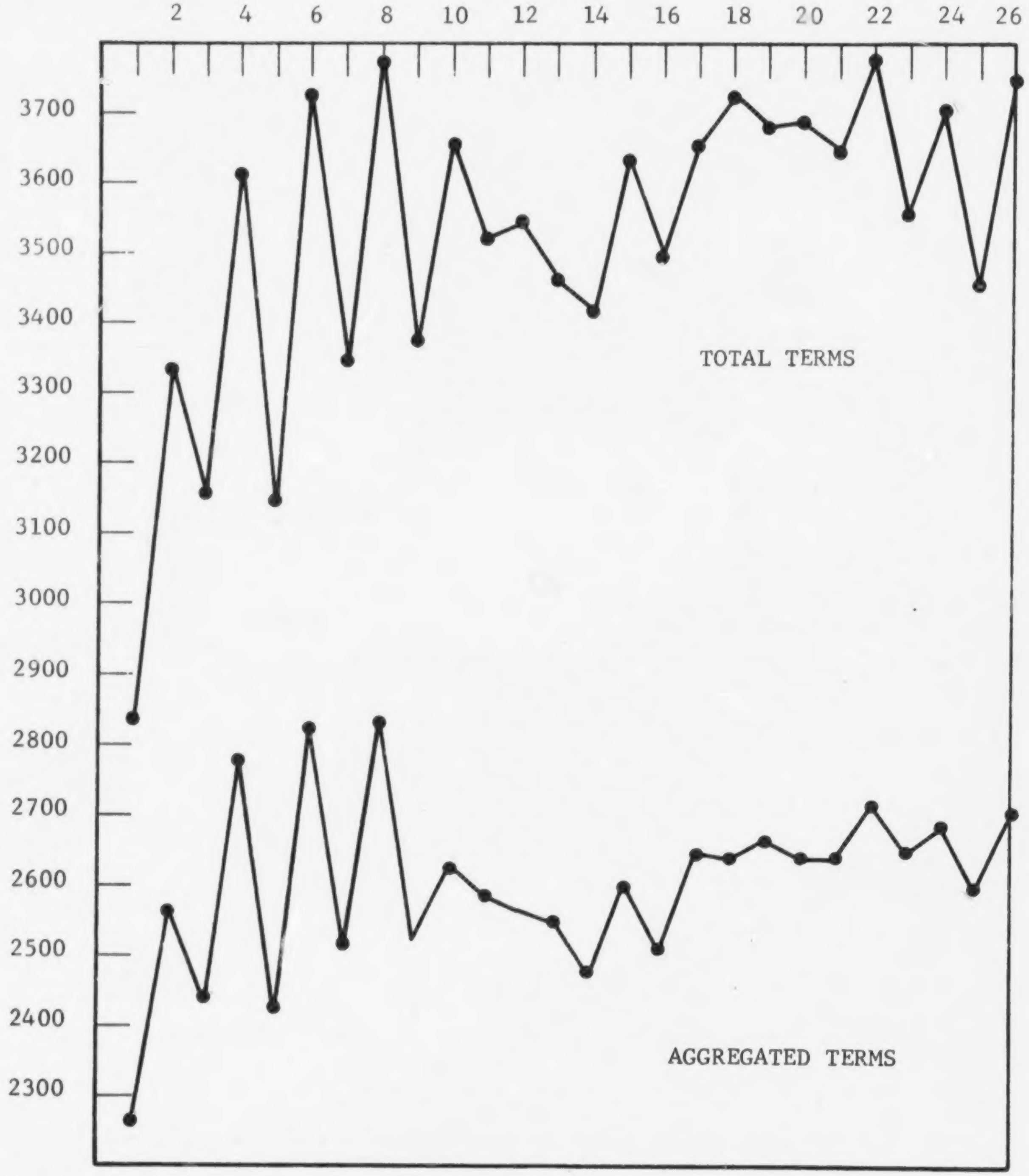


Figure 11-5

TERM AGGREGATION VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

VOLUME 70 VOLUME 71

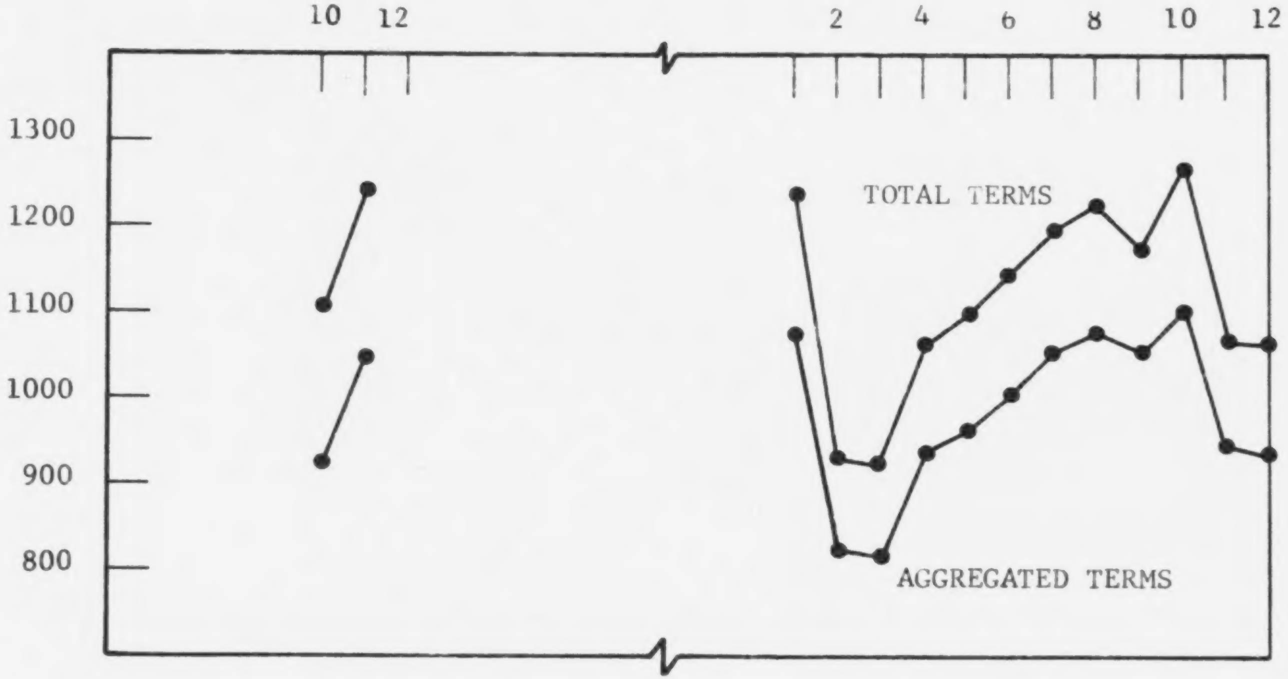


Figure 11-6

TERM AGGREGATION VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

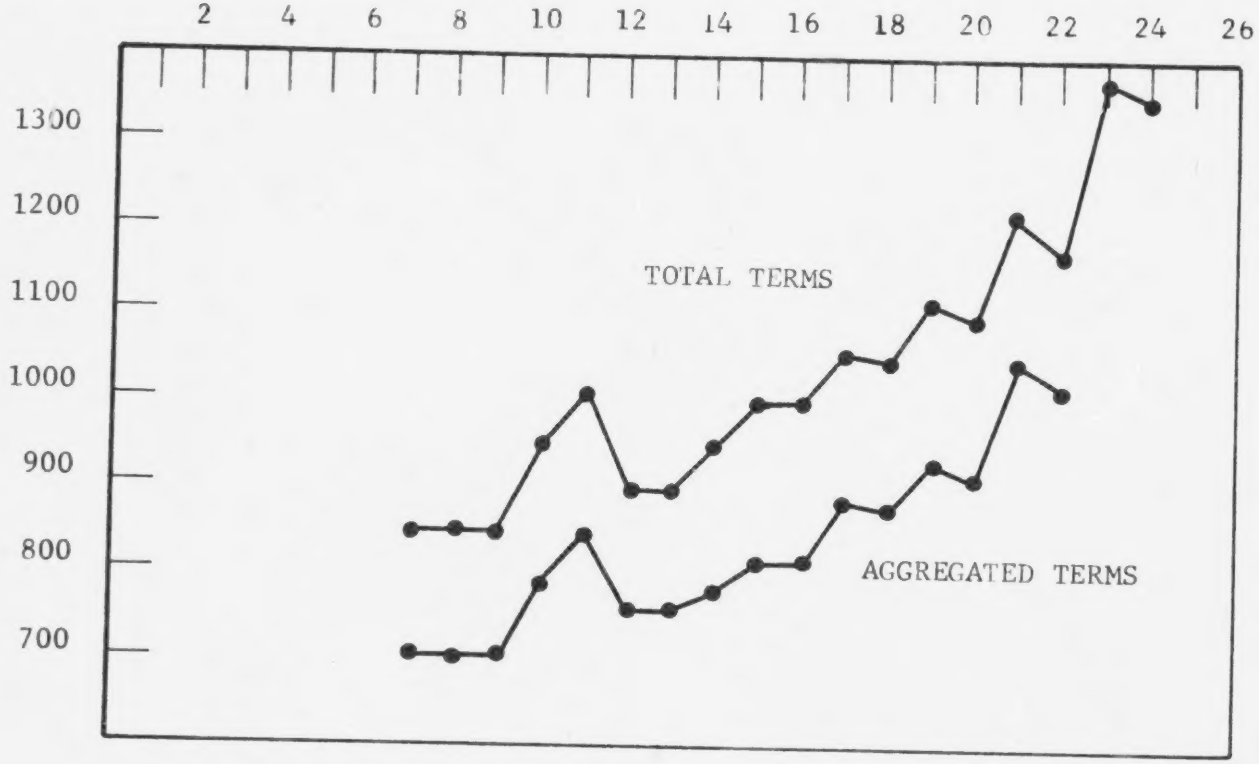


Figure 11-7

TERM AGGREGATION VS. ISSUE

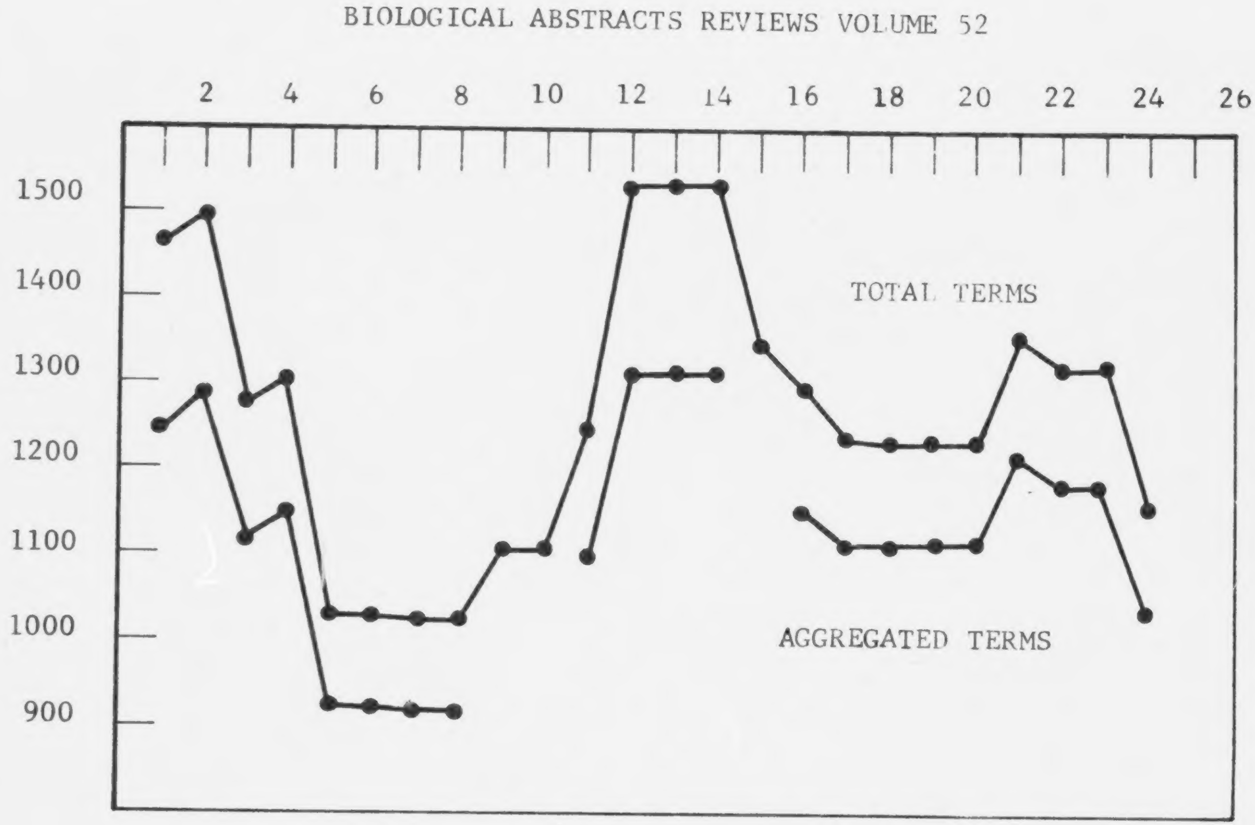


Figure 11-8

TERM AGGREGATION VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

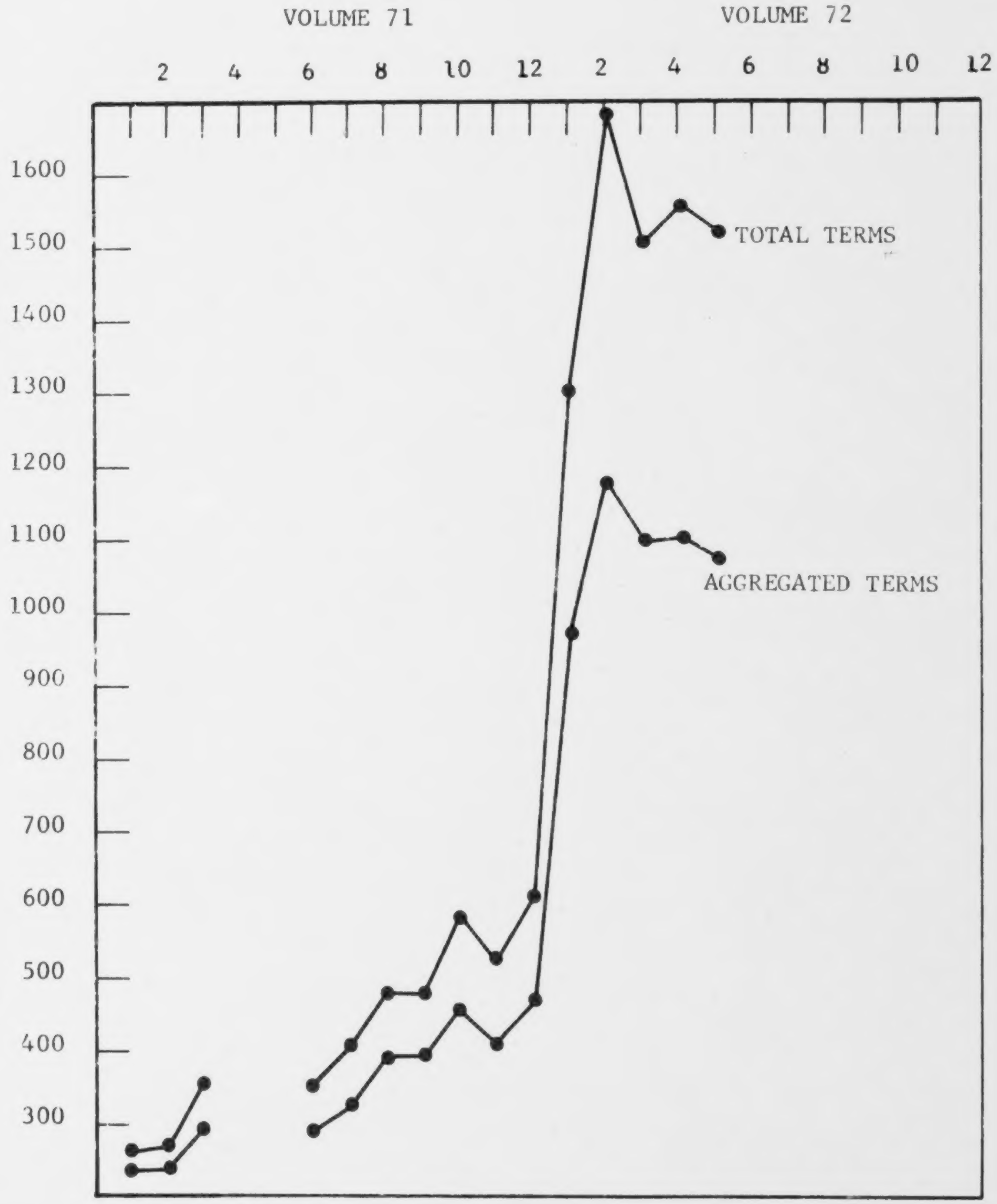


Figure 11-9

TERM AGGREGATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71,72

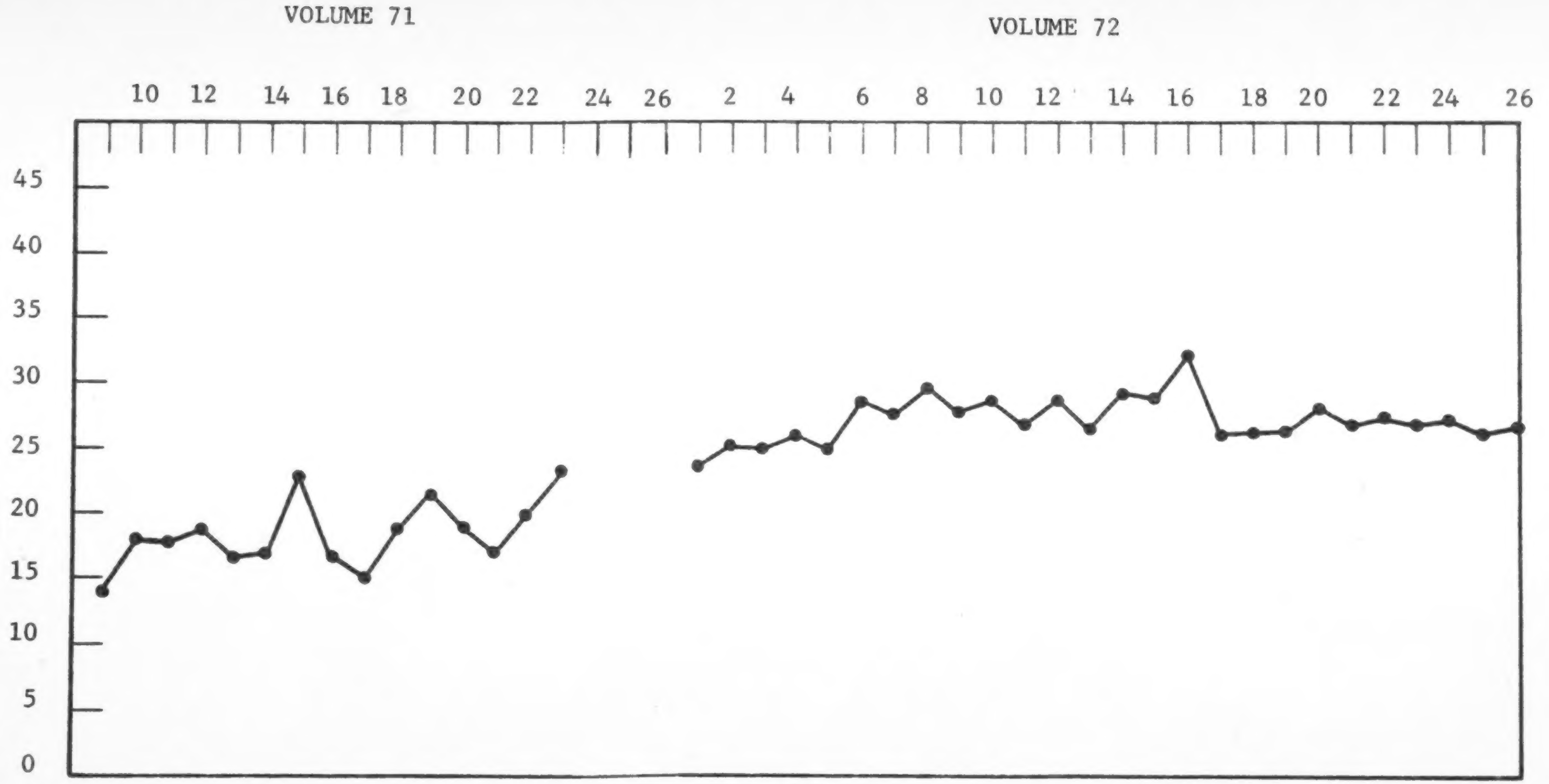


Figure 11-10

AGGREGATION REDUCTION RATIO VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

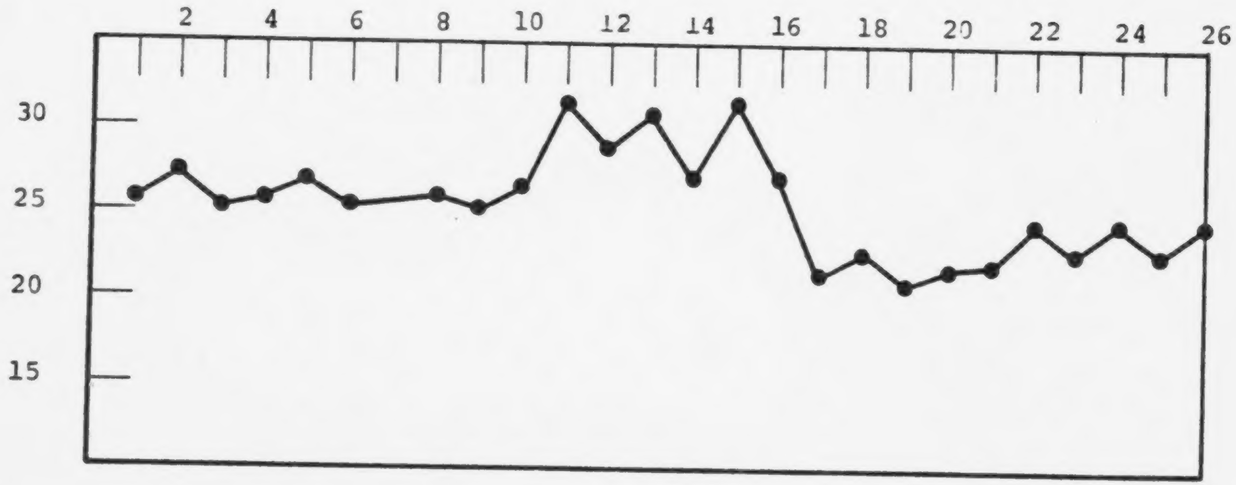


Figure 11-11

AGGREGATION REDUCTION RATIO VS. ISSUE

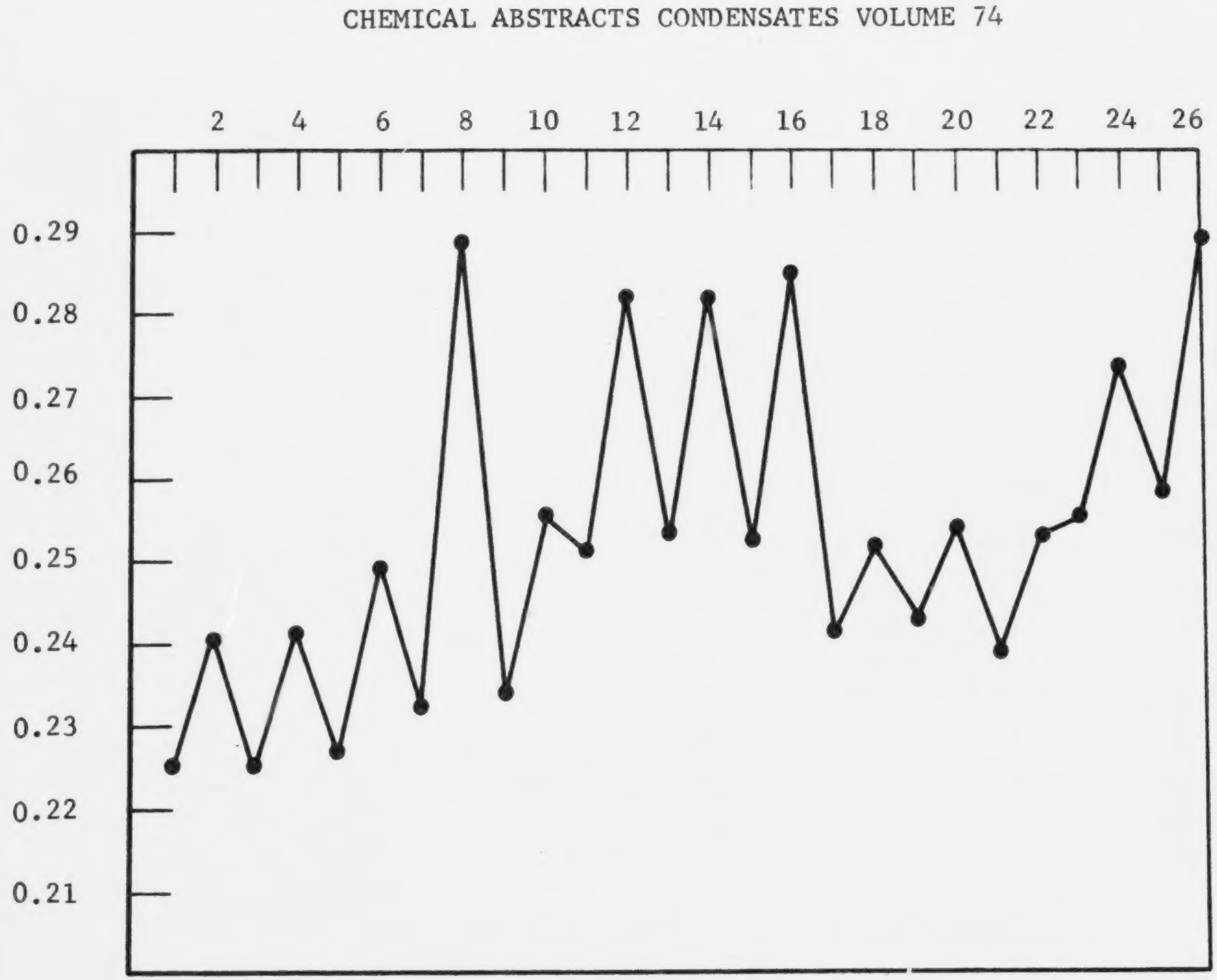


Figure 11-12

AGGREGATION REDUCTION RATIO VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

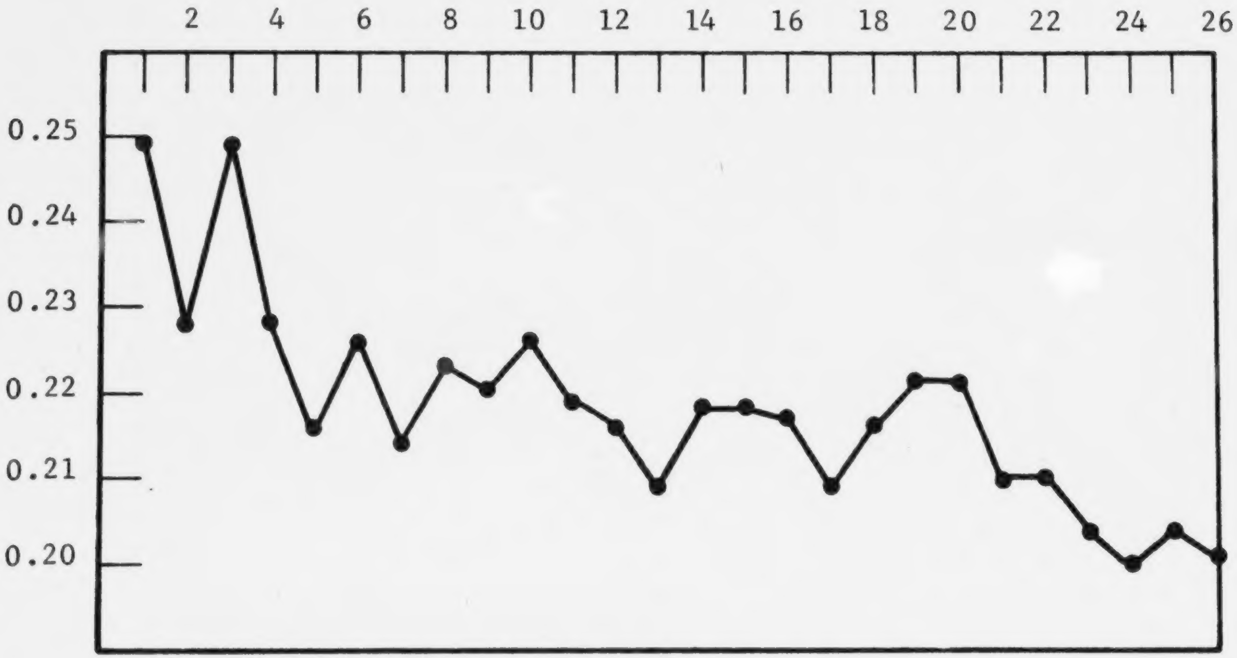


Figure 11-13

AGGREGATION REDUCTION RATIO VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

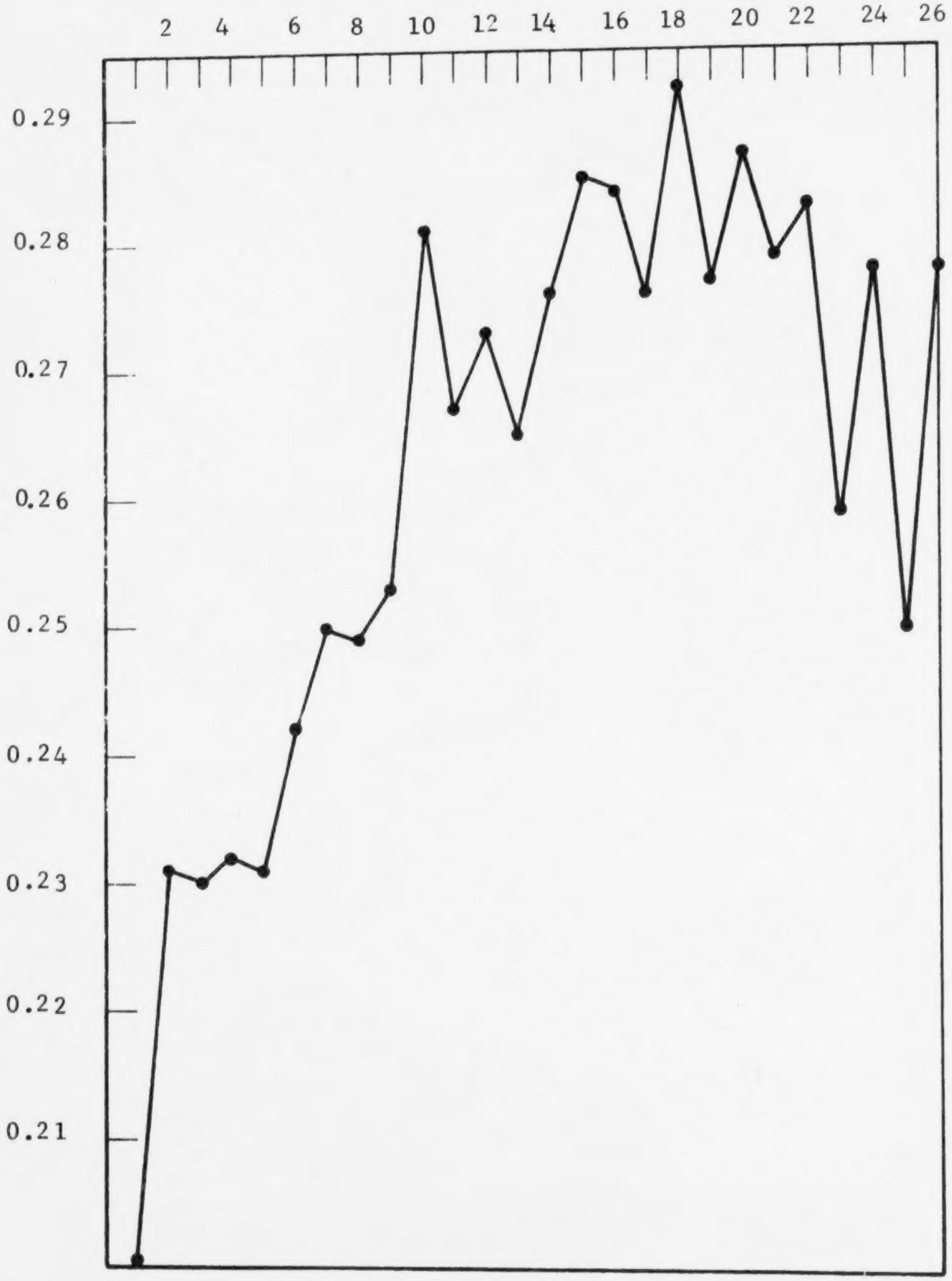


Figure 11-14
AGGREGATION REDUCTION RATIO VS. ISSUE

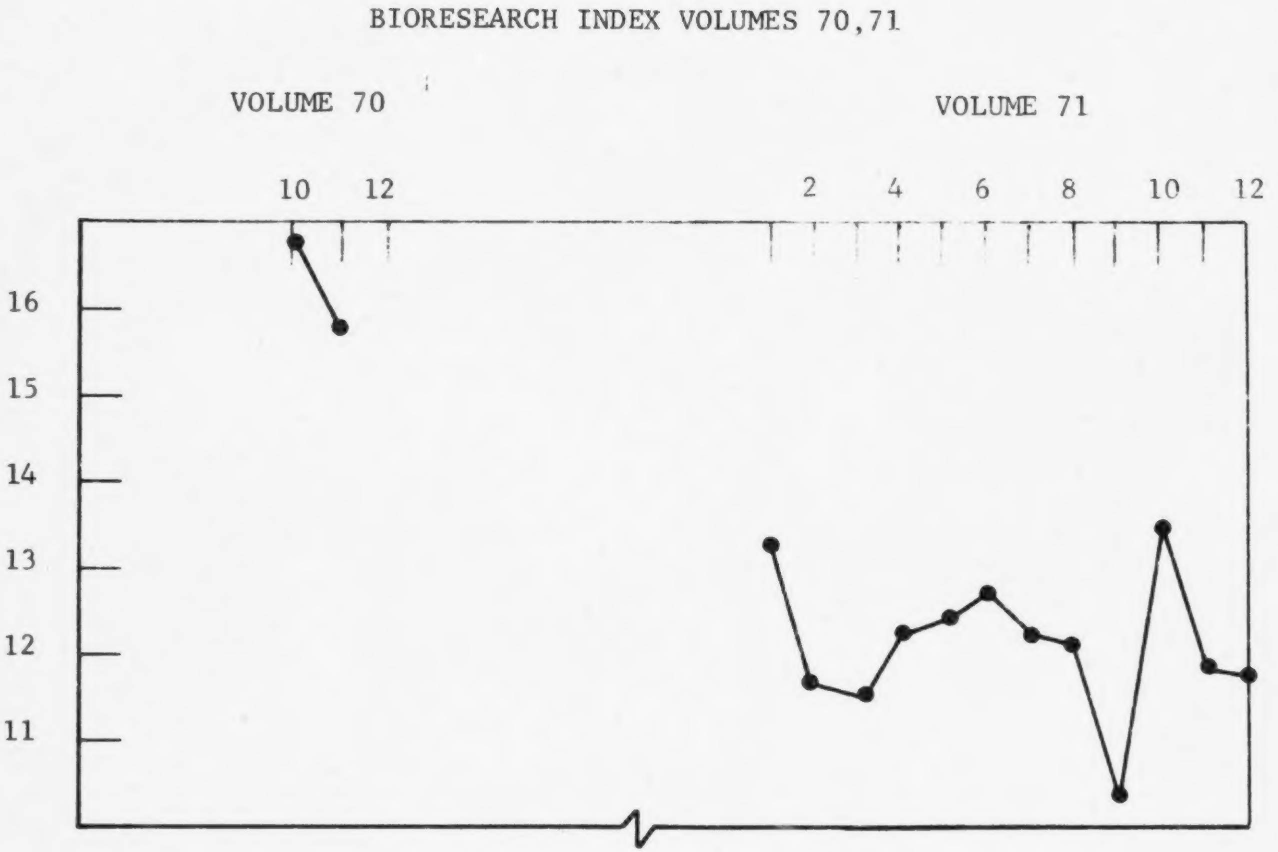


Figure 11-15

AGGREGATION REDUCTION RATIO VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

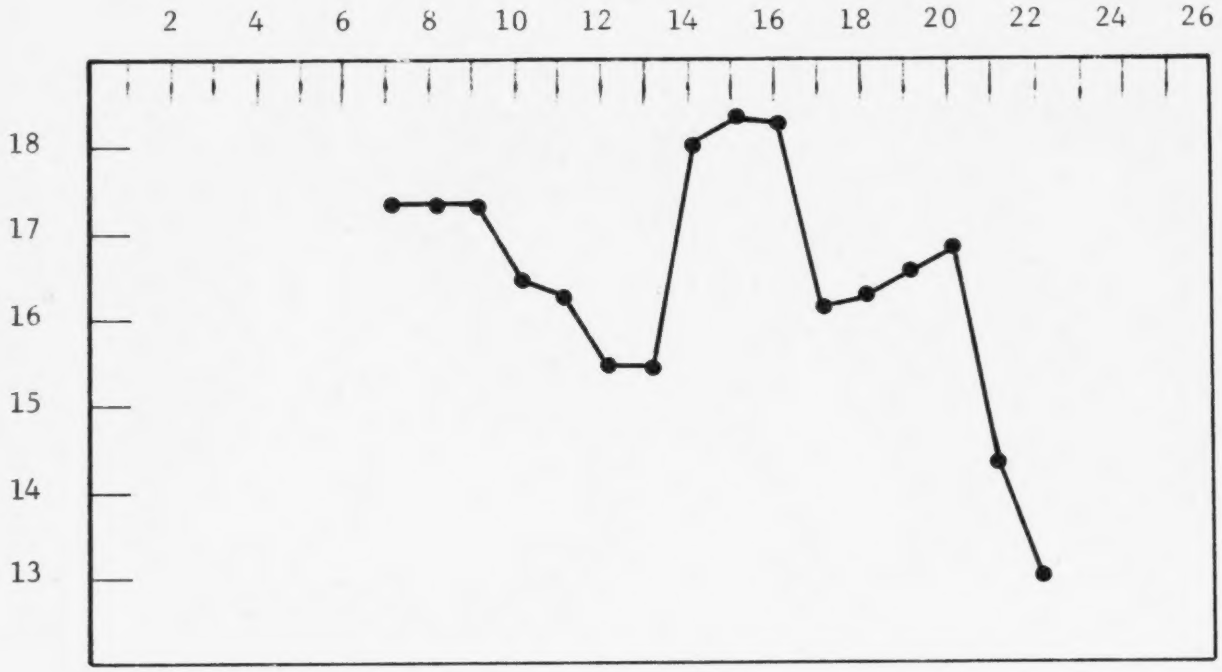


Figure 11-16

AGGREGATION REDUCTION RATIO VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

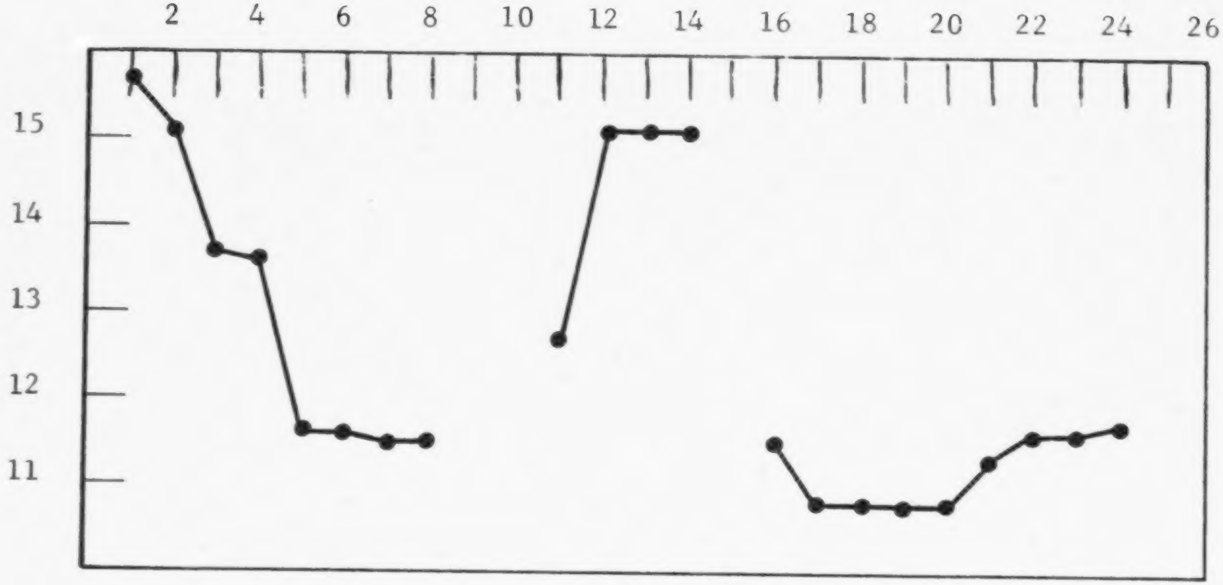


Figure 11-17

AGGREGATION REDUCTION RATIO VS. ISSUE

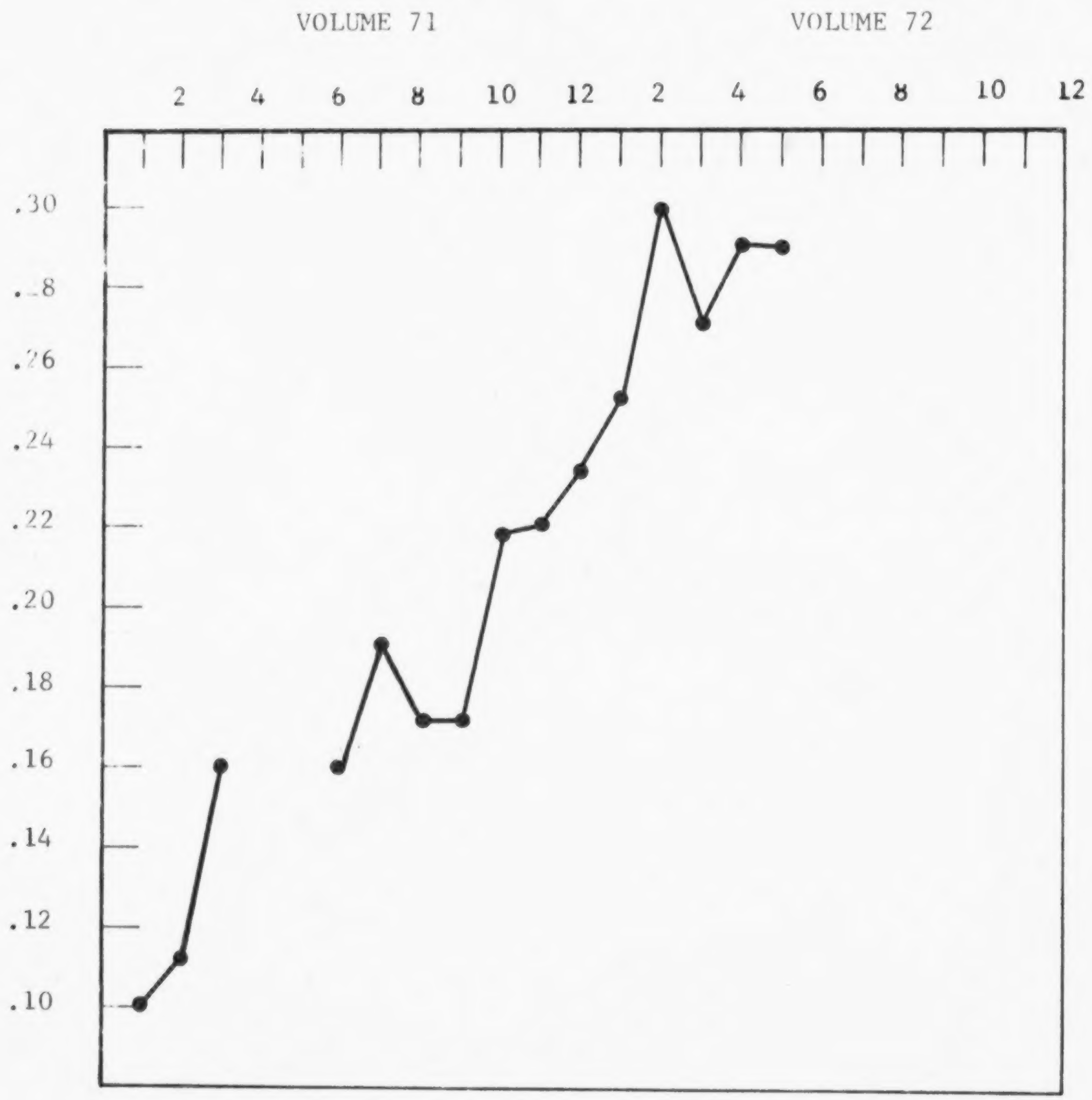

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

Figure 11-18

AGGREGATION REDUCTION RATIO VS. ISSUE

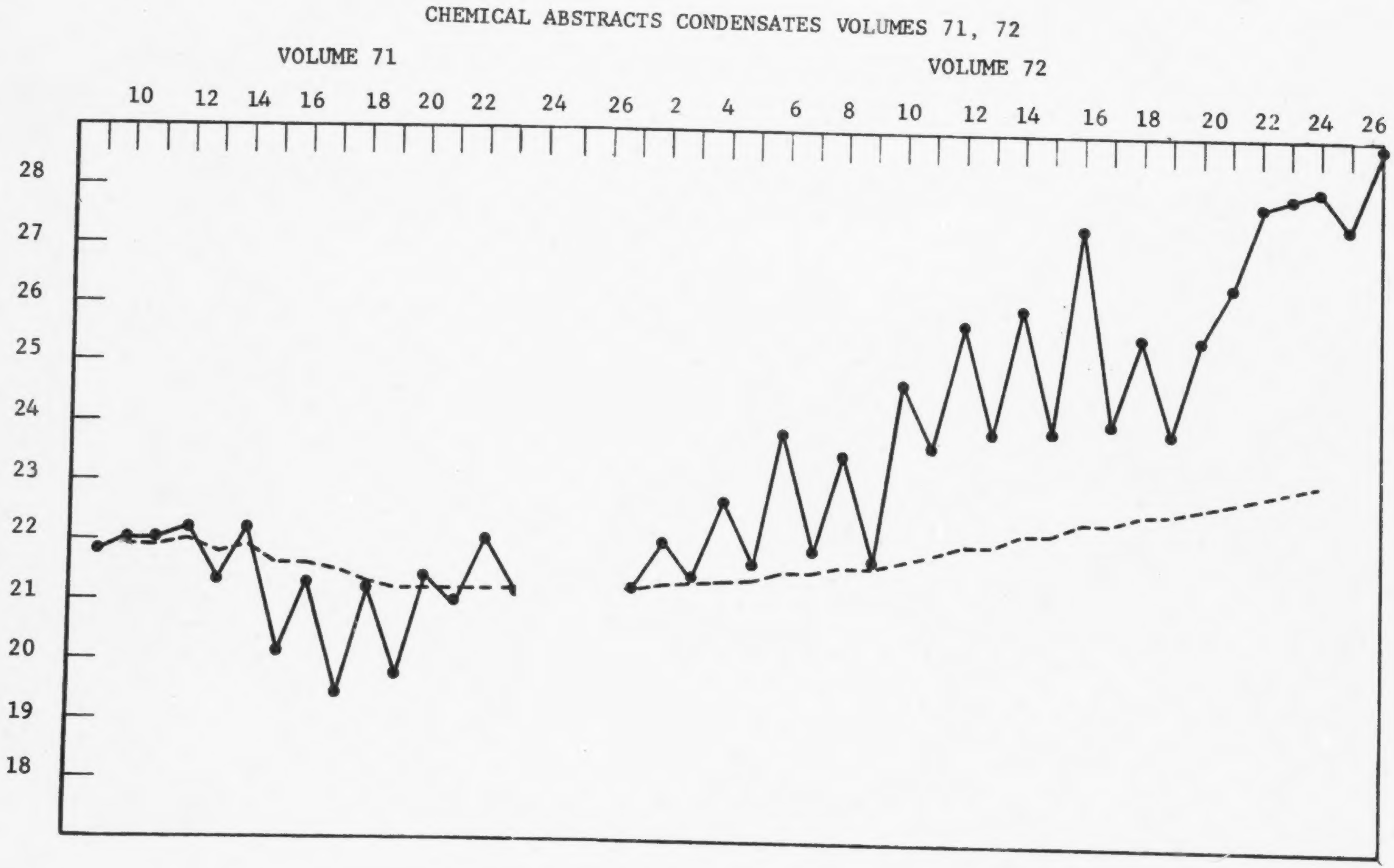


Figure 11-19

TERMS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

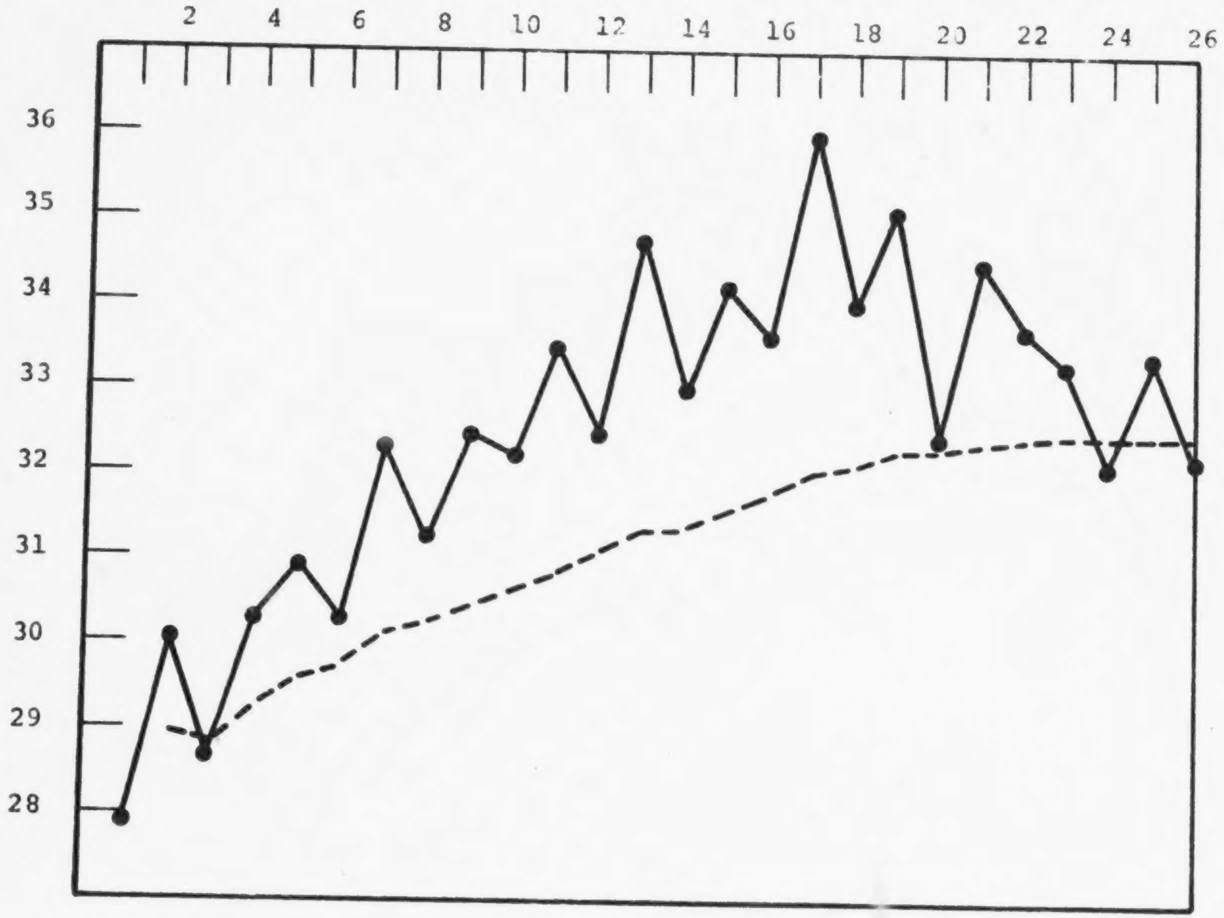


Figure 11-20
TERMS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

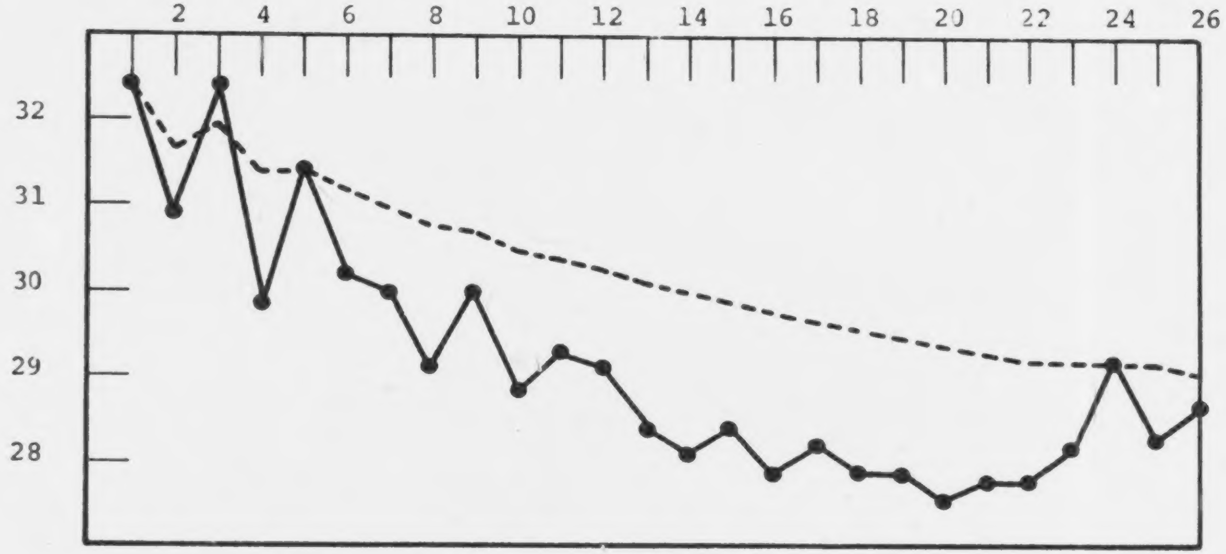


Figure 11-21

TERMS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

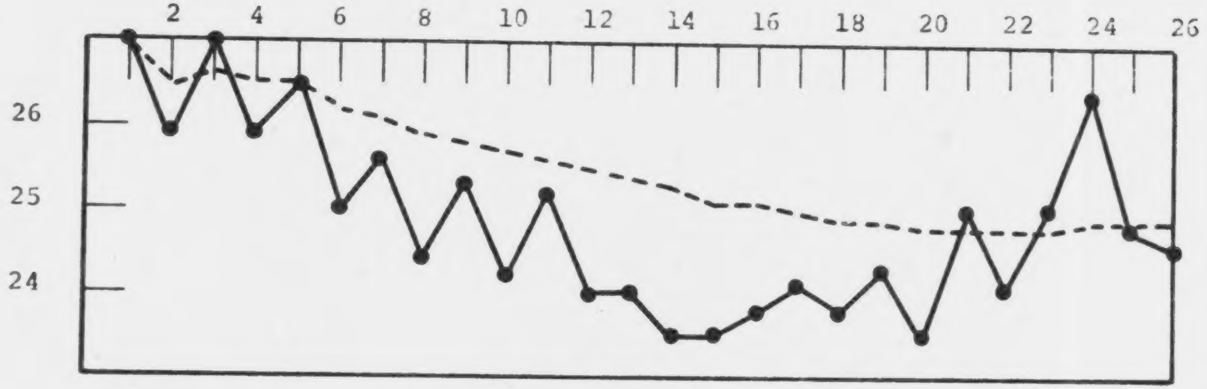


Figure 11-22

TERMS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

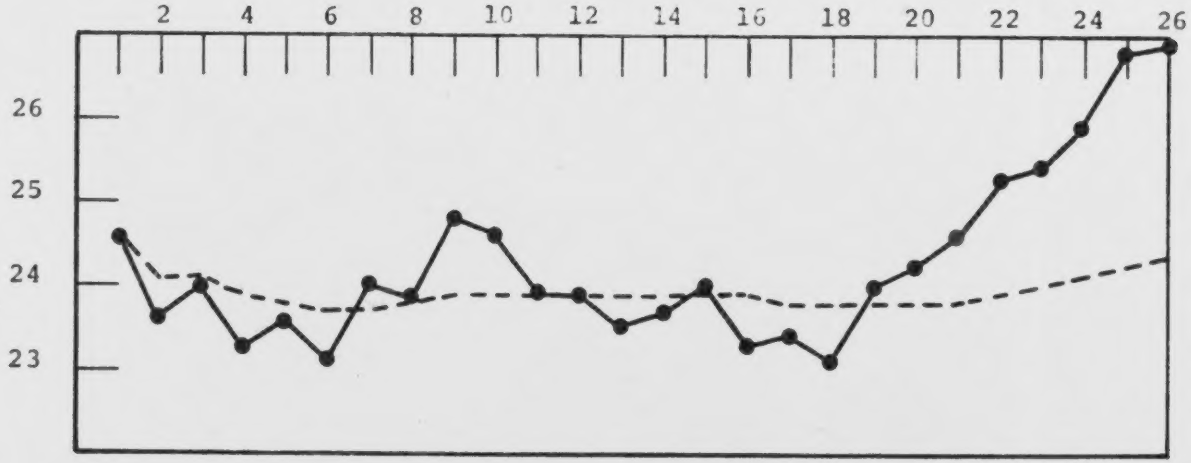


Figure 11-23

TERMS PER PROFILE VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

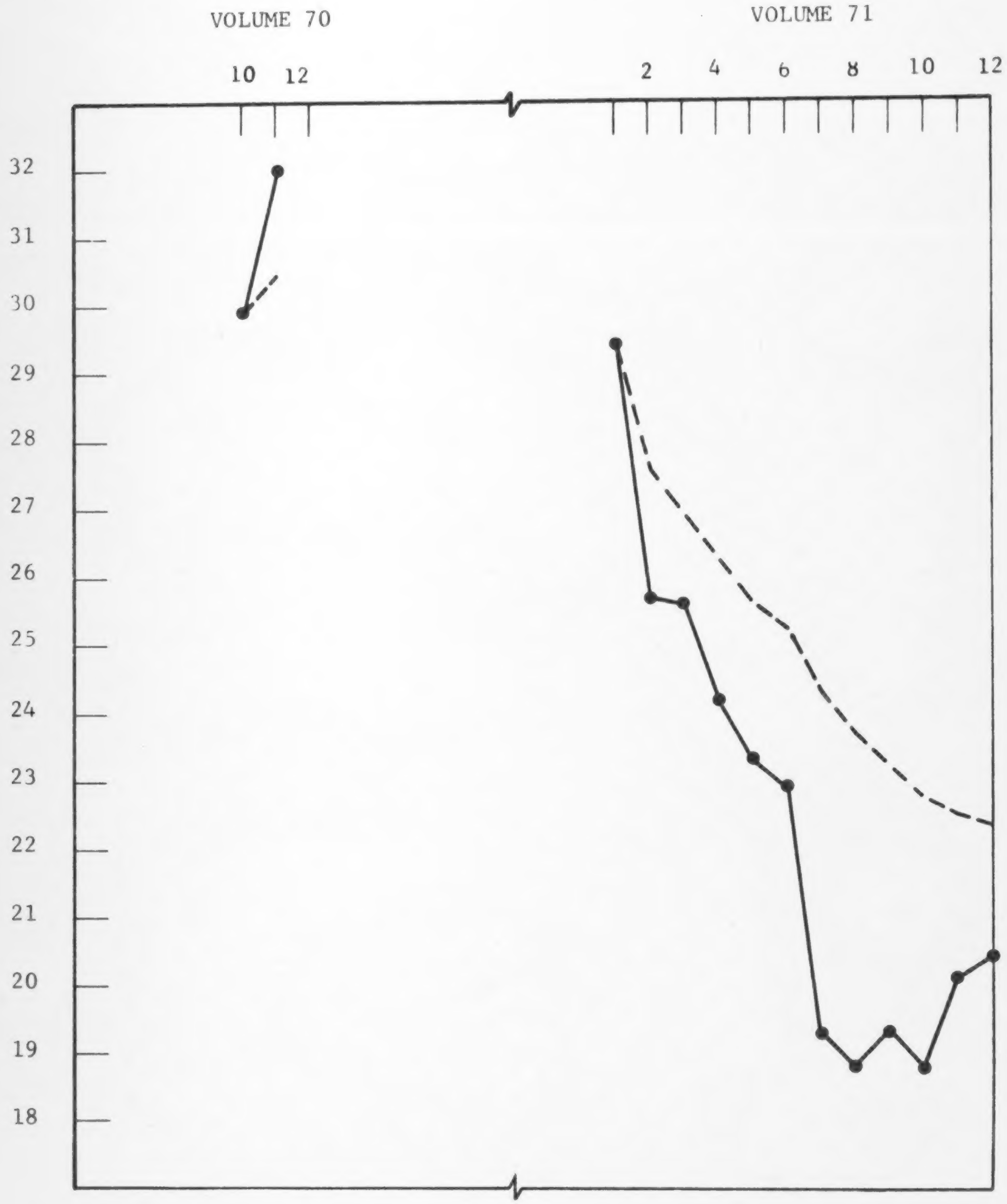


Figure 11-24

TERMS PER PROFILE VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

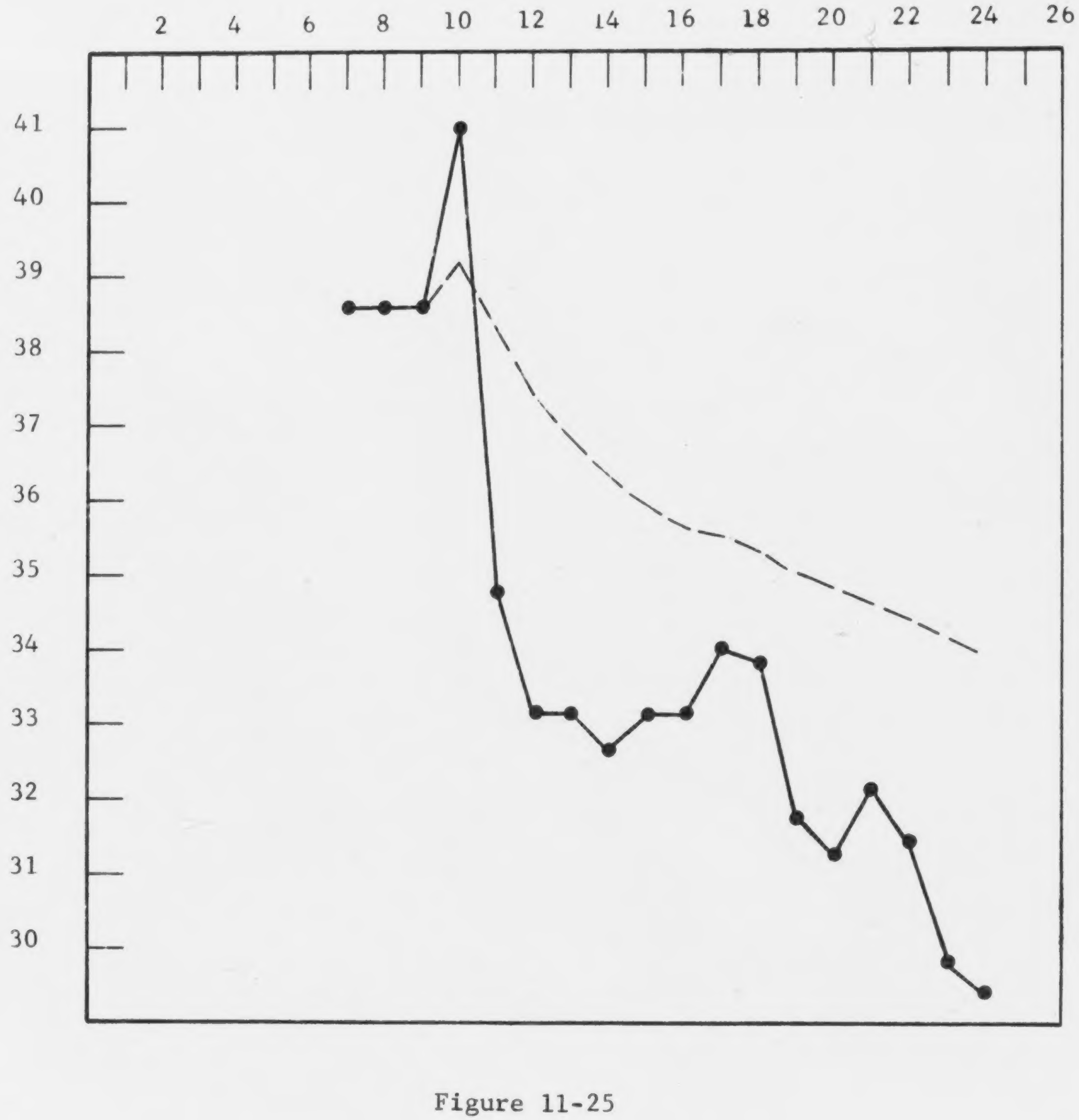


Figure 11-25

TERMS PER PROFILE VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

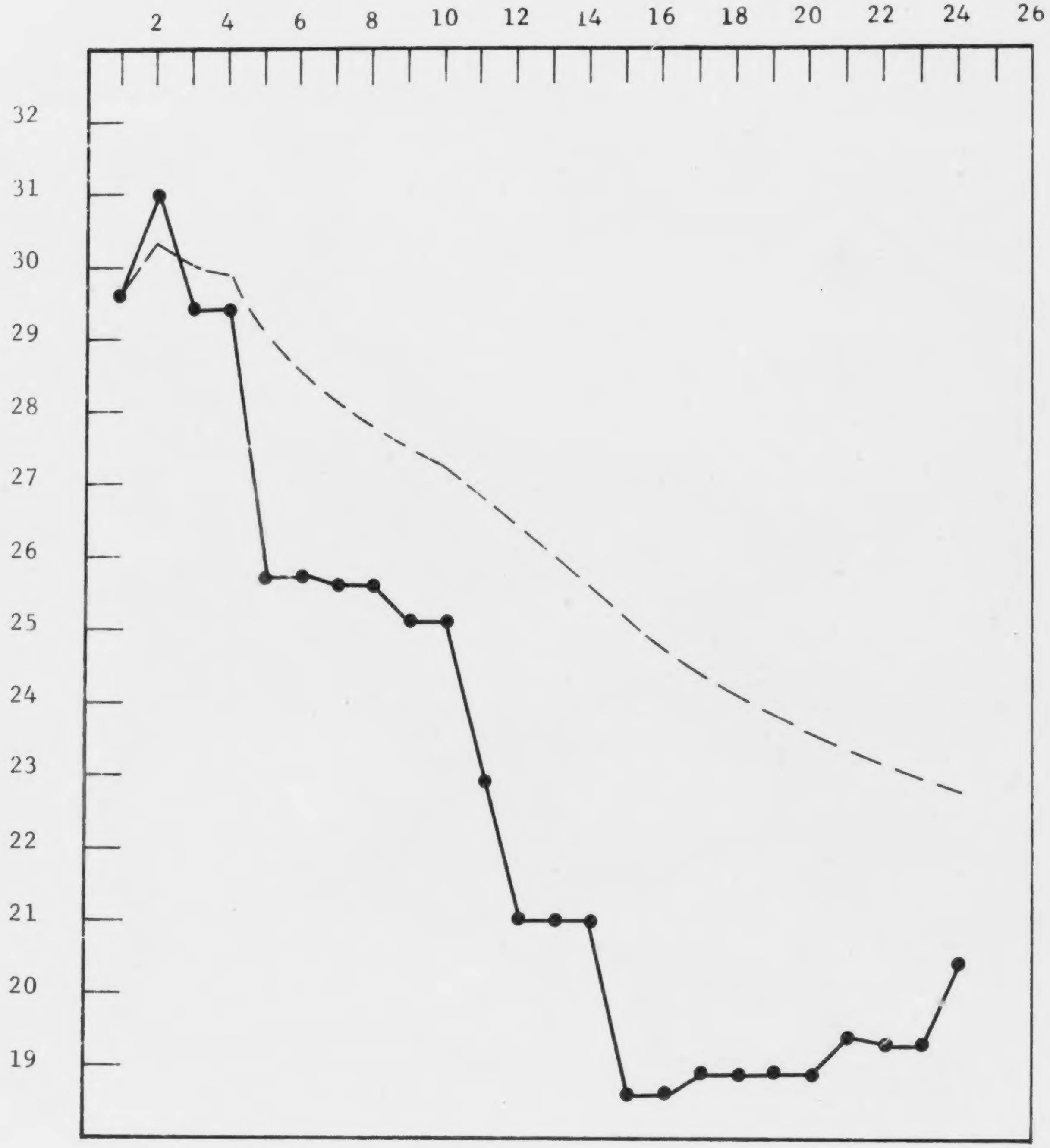


Figure 11-26

TERMS PER PROFILE VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

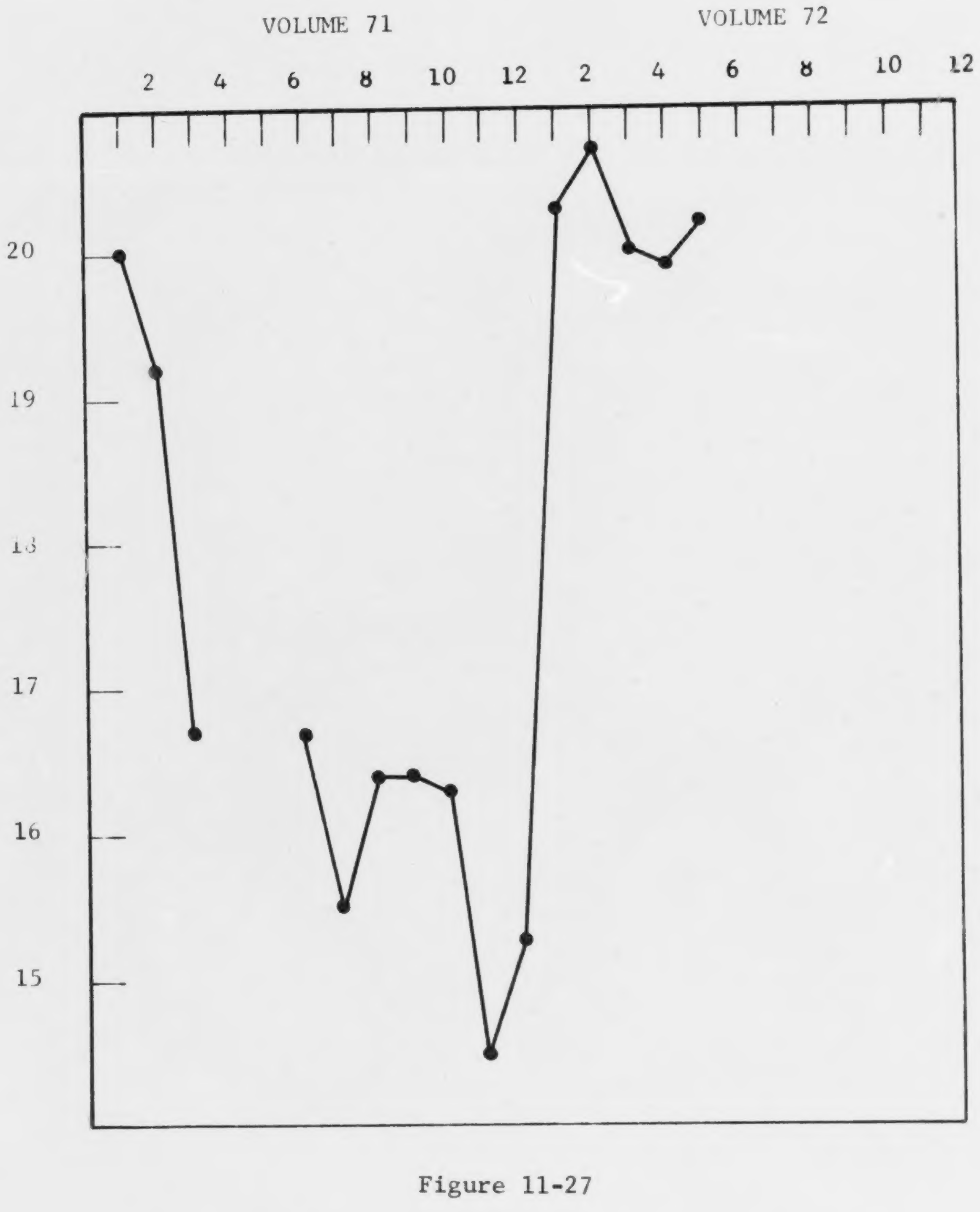


Figure 11-27

TERMS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71, 72

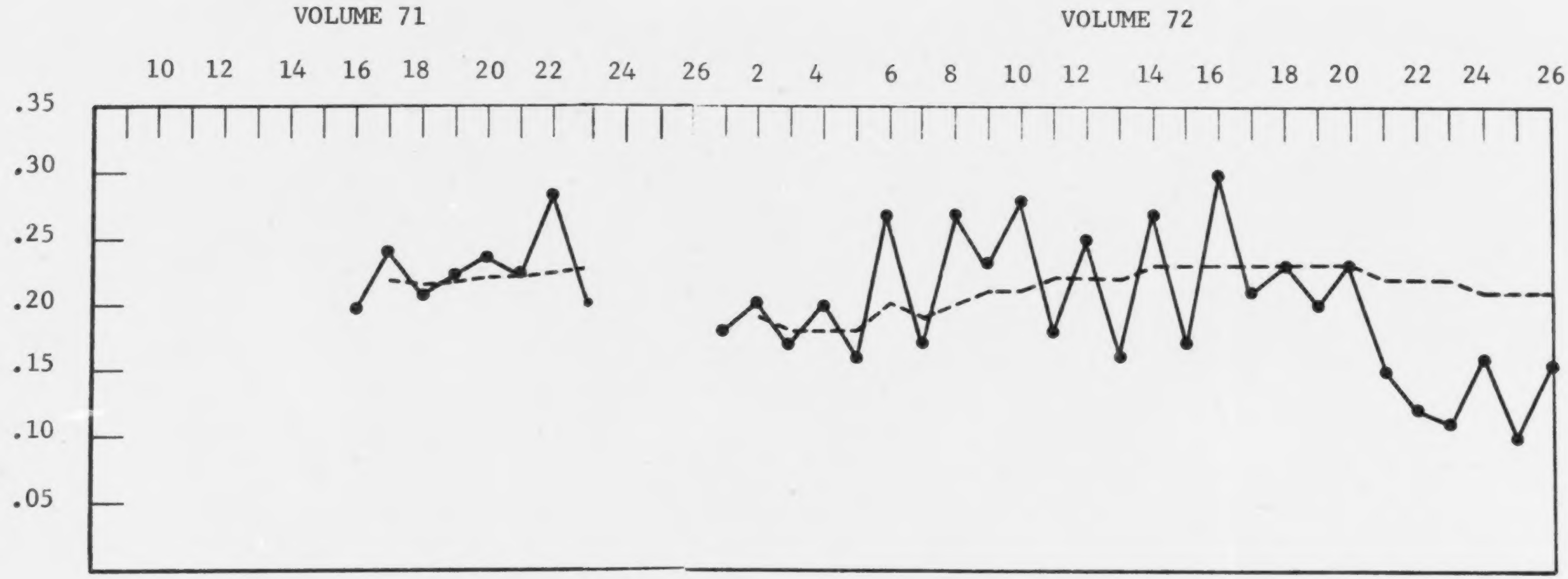


Figure 11-28

COST PER TERM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

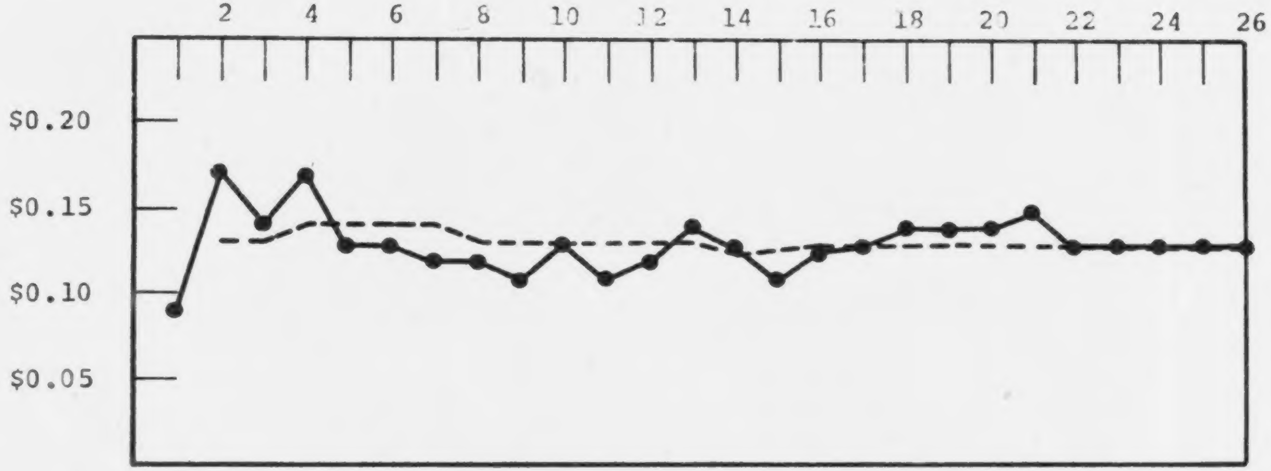


Figure 11-29

COST PER TERM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

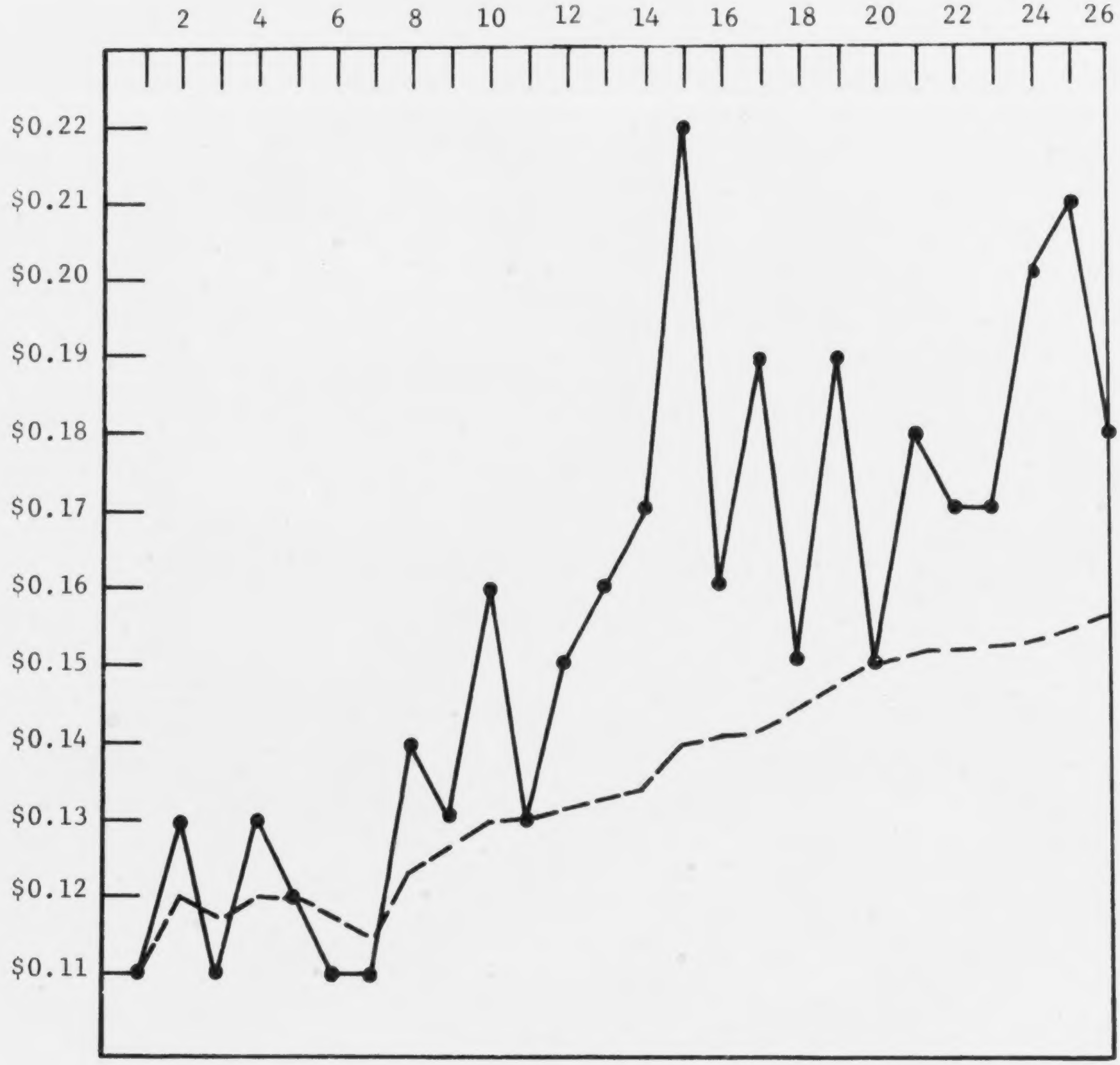


Figure 11-30

COST PER TERM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

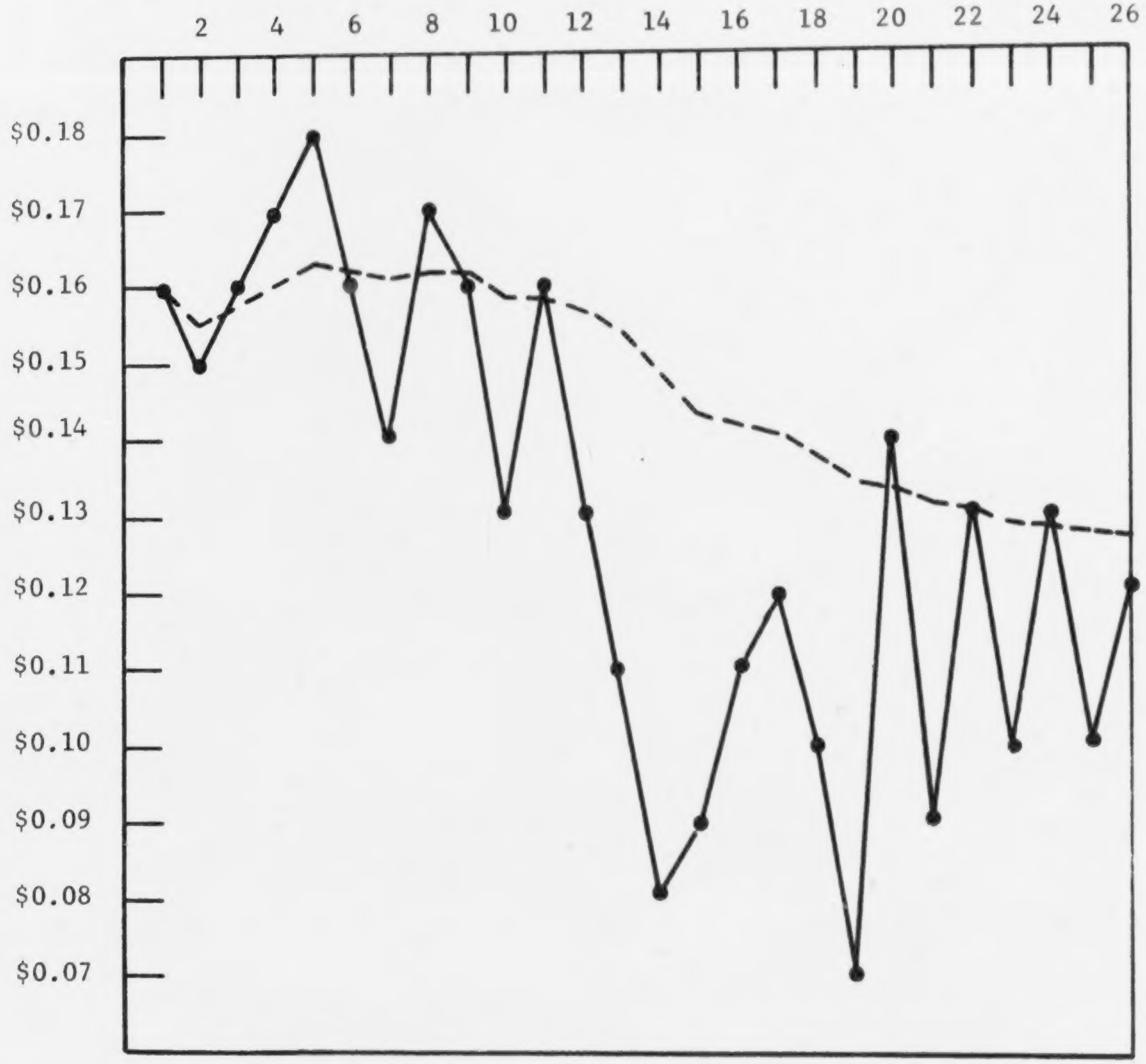


Figure 11-31

COST PER TERM VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

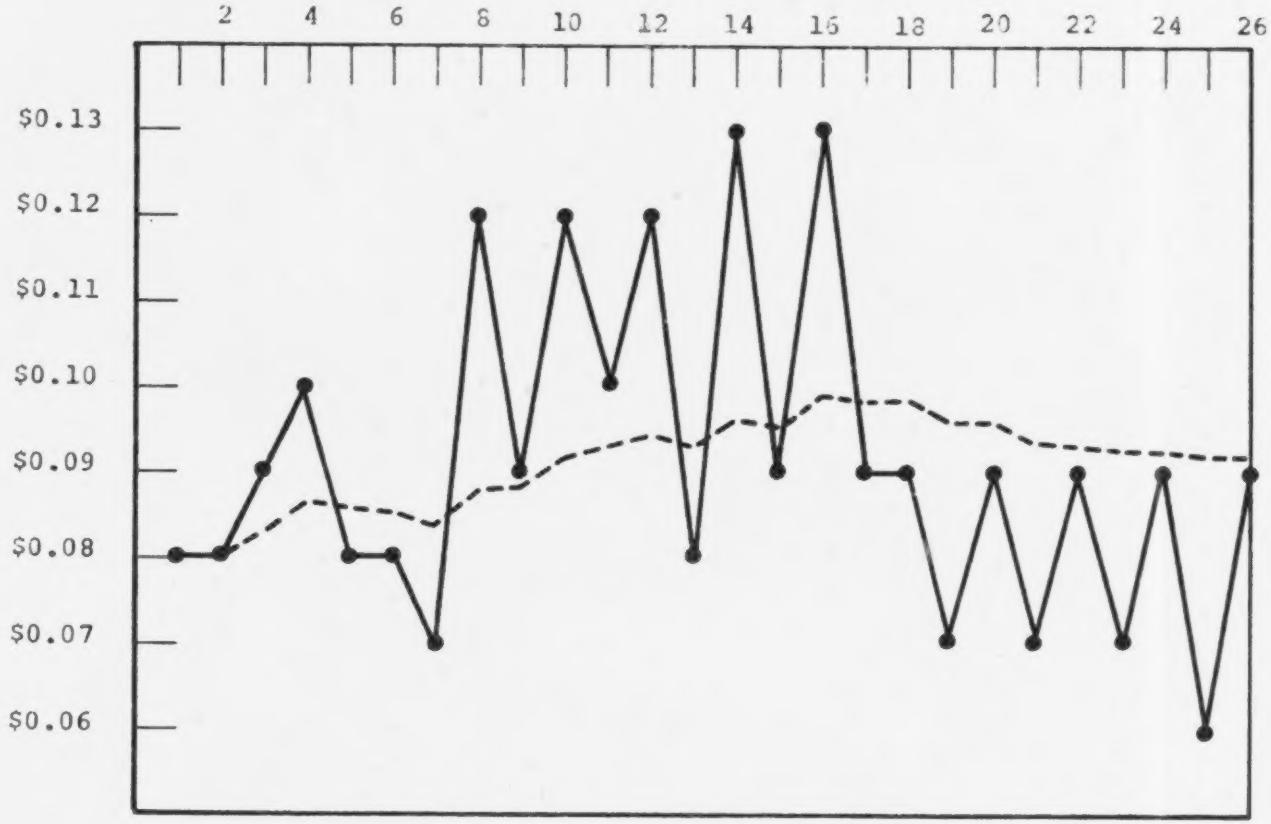


Figure 11-32

COST PER TERM VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

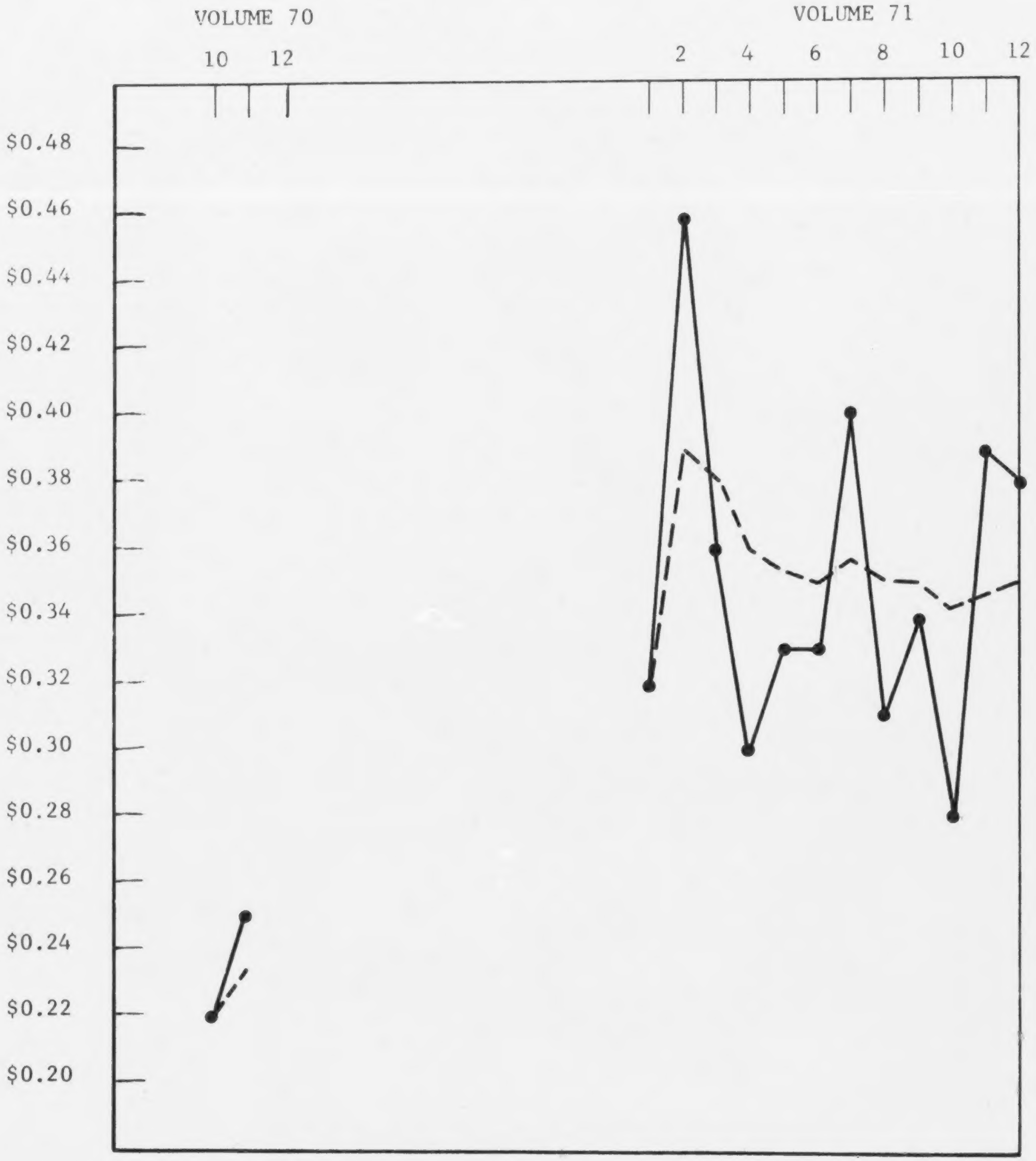


Figure 11-33

COST PER TERM VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

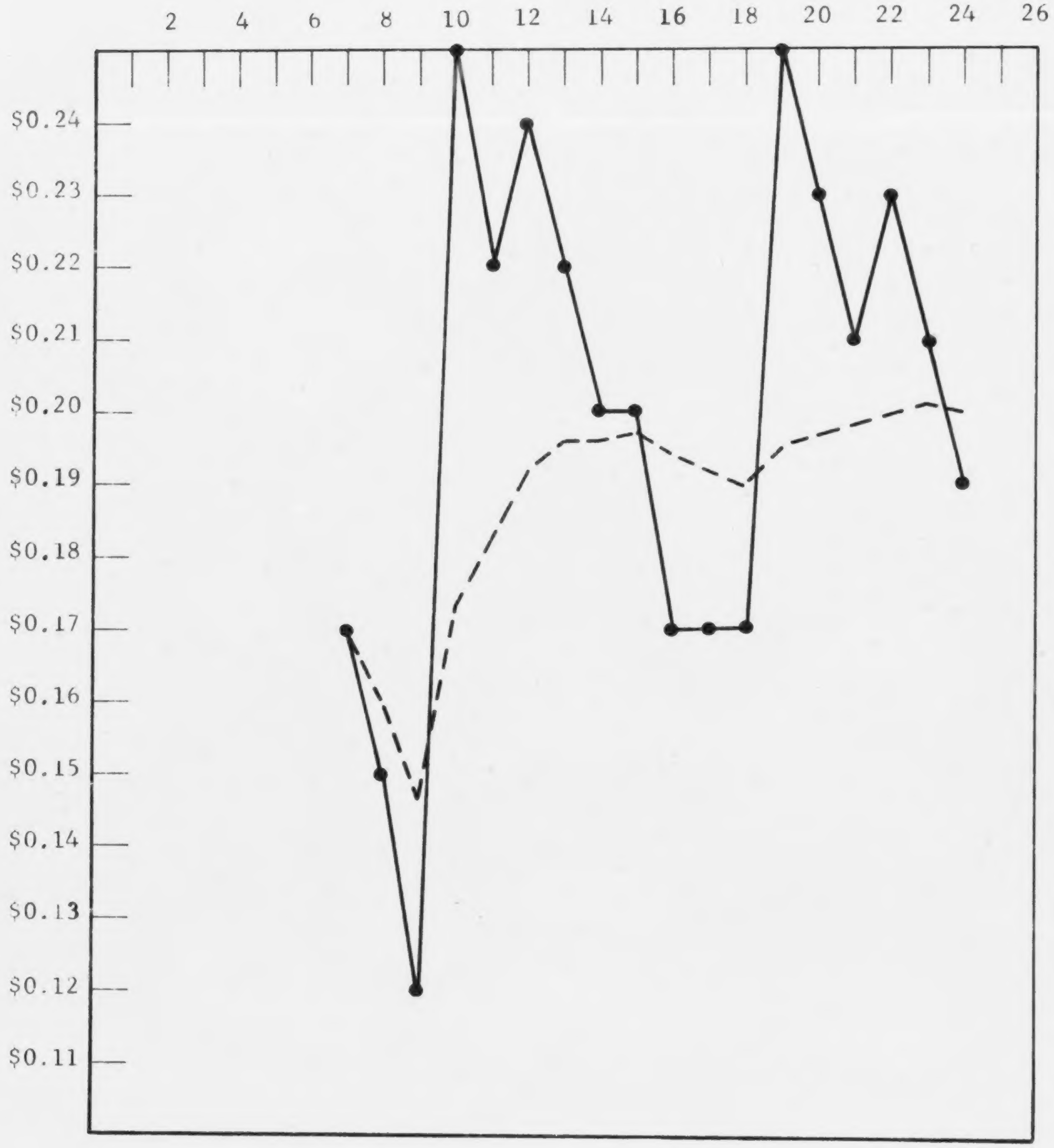


Figure 11-34

COST PER TERM VS. ISSUE

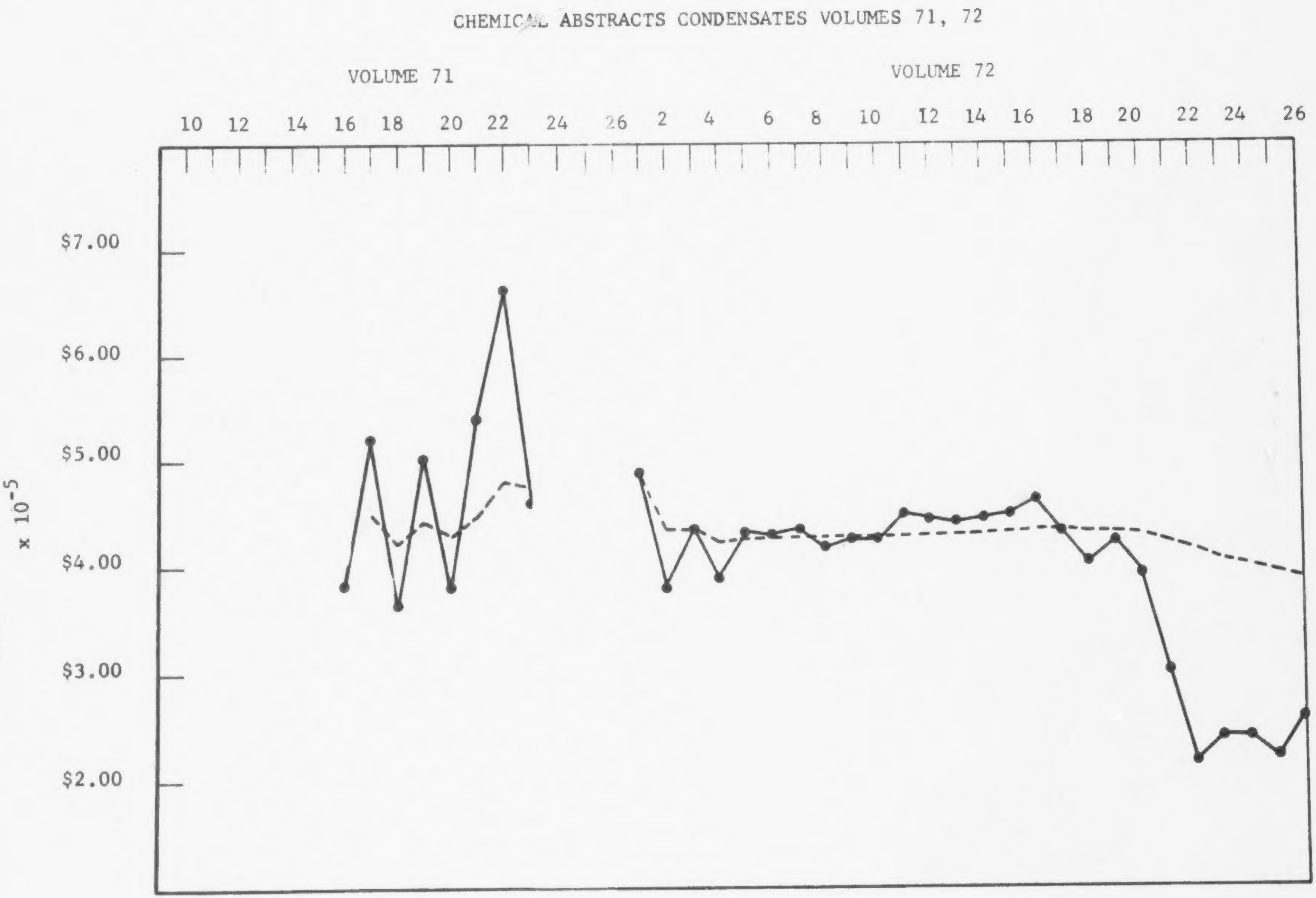


Figure 11-37

COST PER TERM PER CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

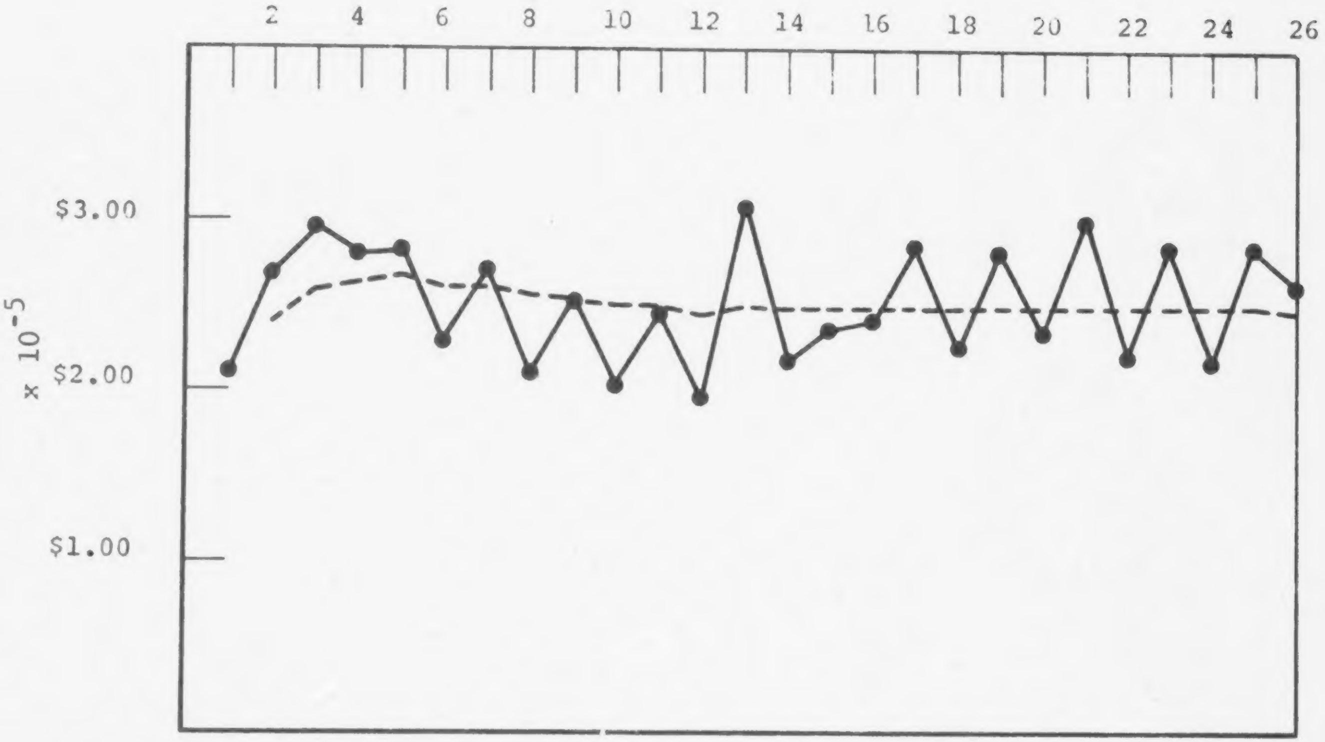


Figure 11-38

COST PER TERM PER CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

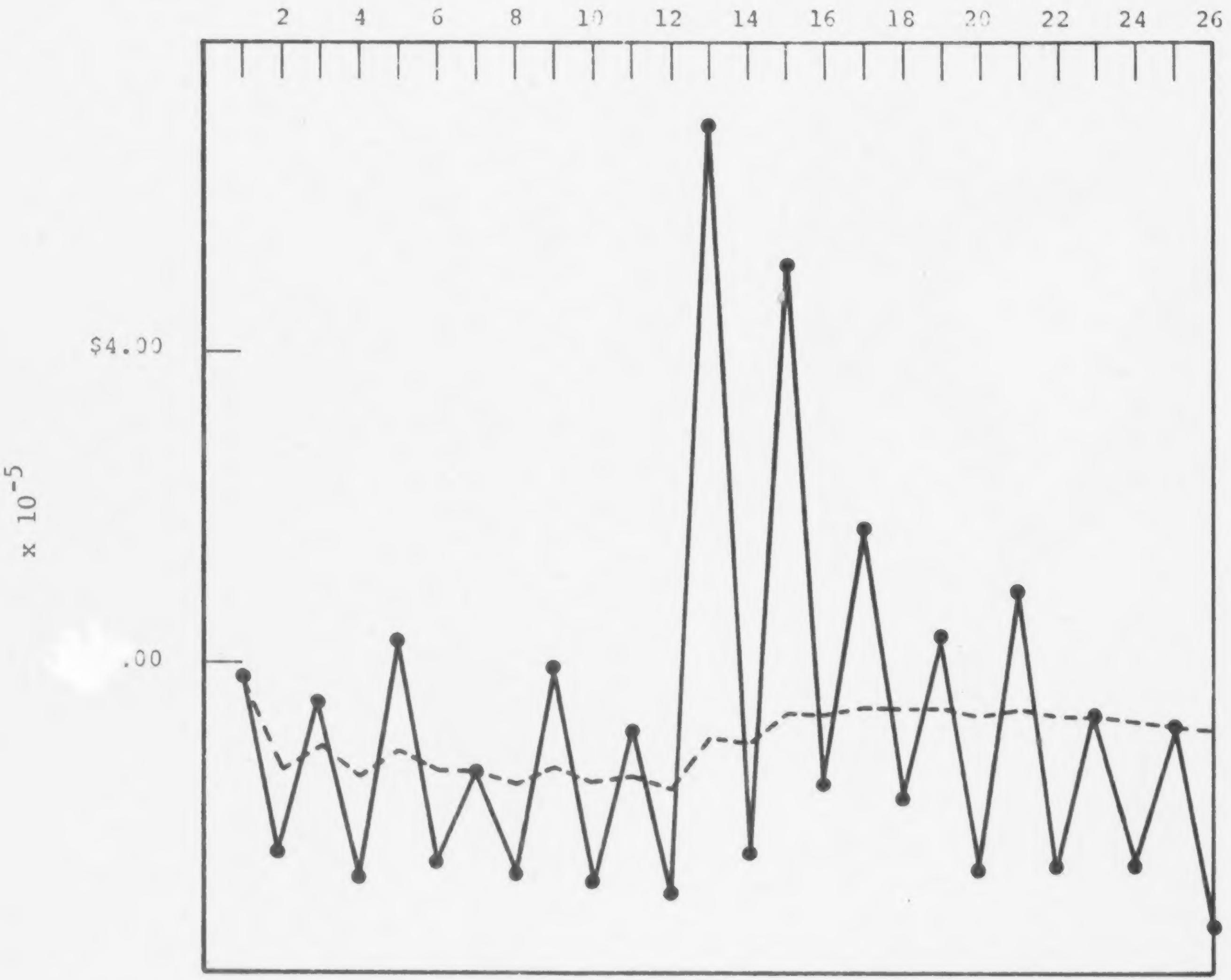


Figure 11-39

COST PER TERM PER CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

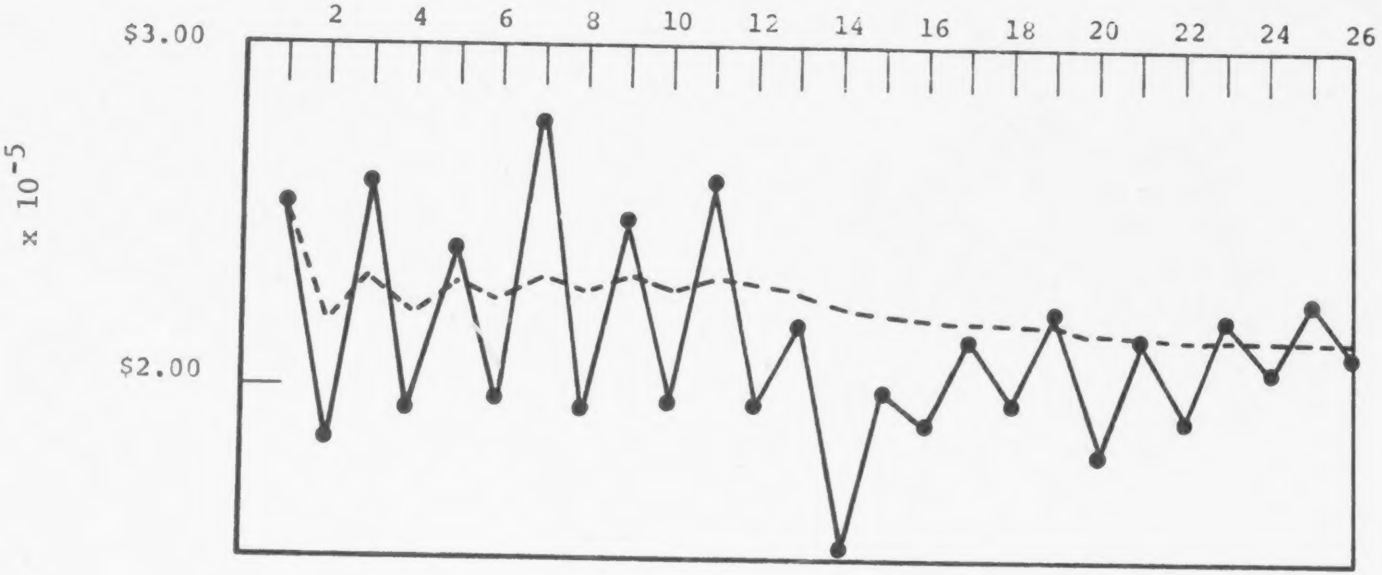


Figure 11-40

COST PER TERM PER CITATION VS. ISSUE

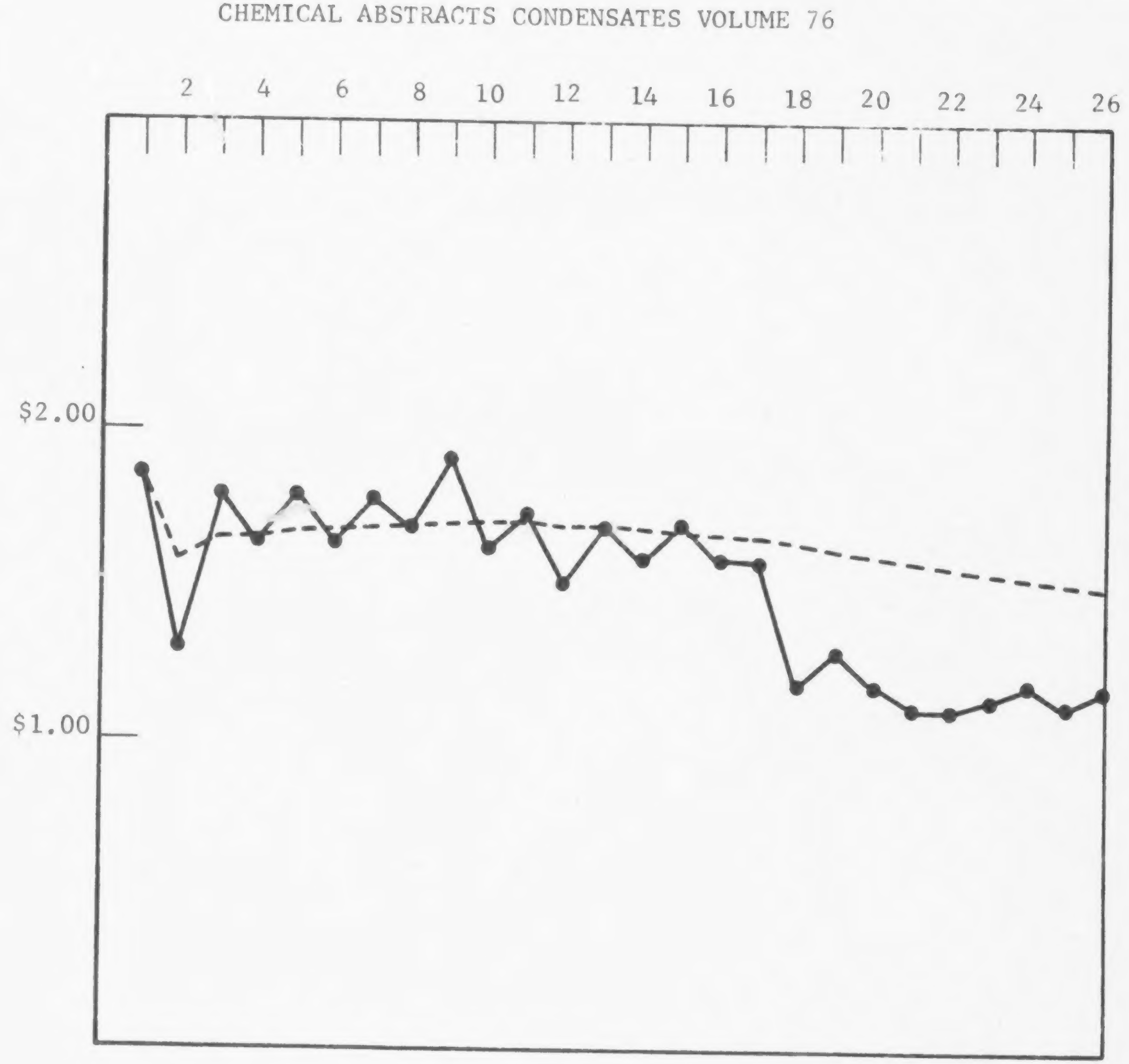


Figure 11-41

COST PER TERM PER CITATION VS. ISSUE

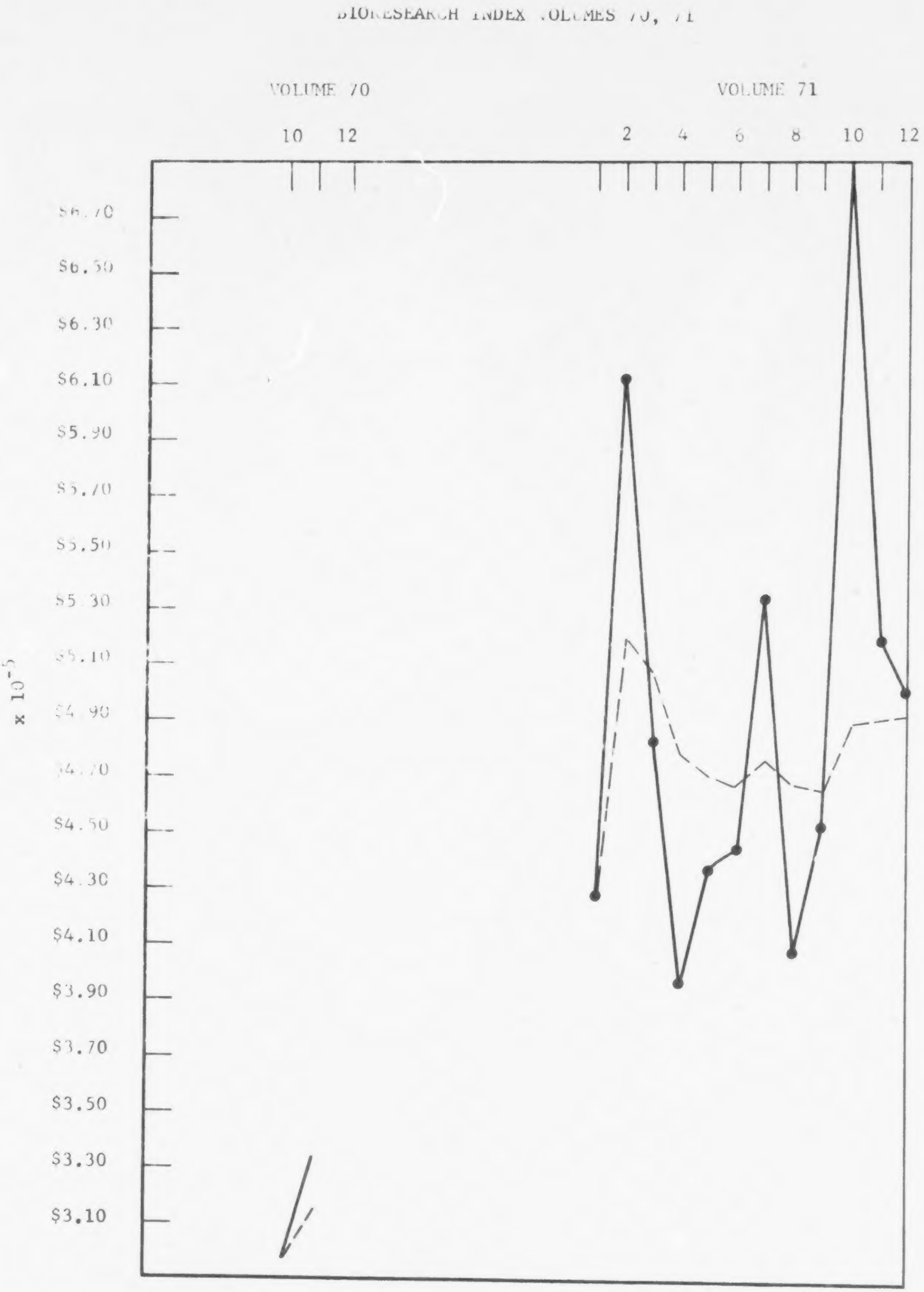


Figure 11-42

COST PER TERM PER CITATION VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

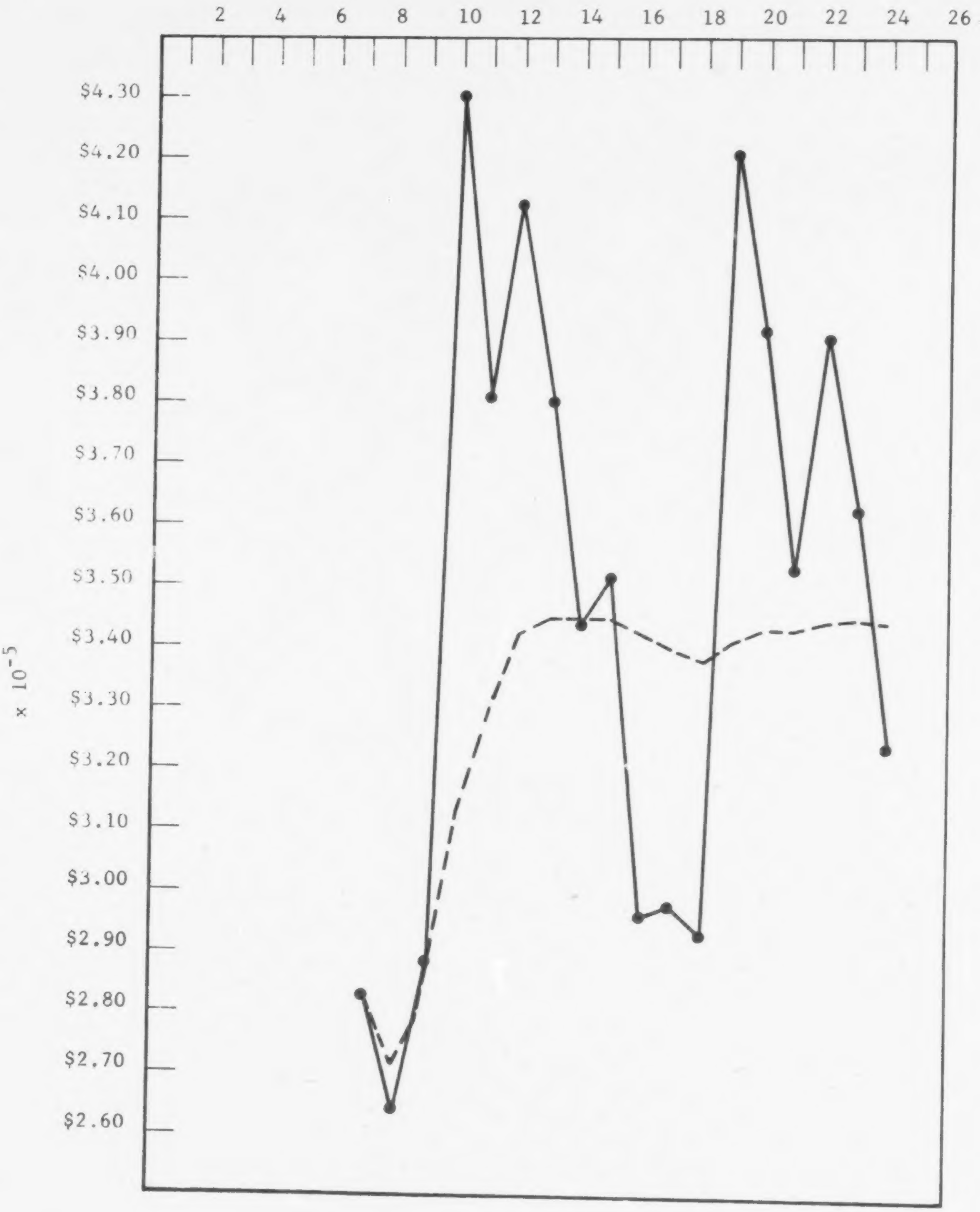


Figure 11-43

COST PER TERM PER CITATION VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

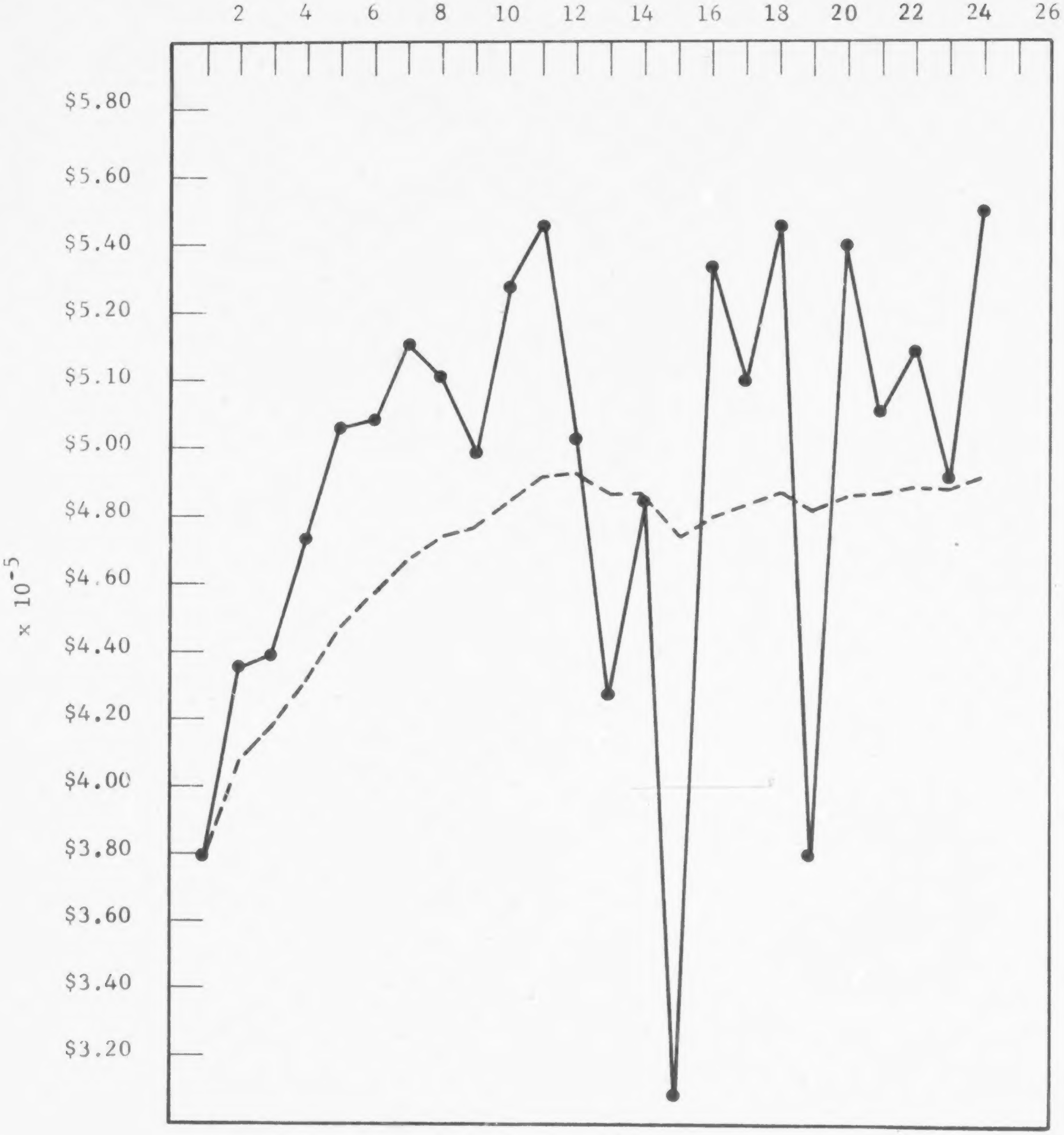


Figure 11-44

COST PER TERM PER CITATION VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

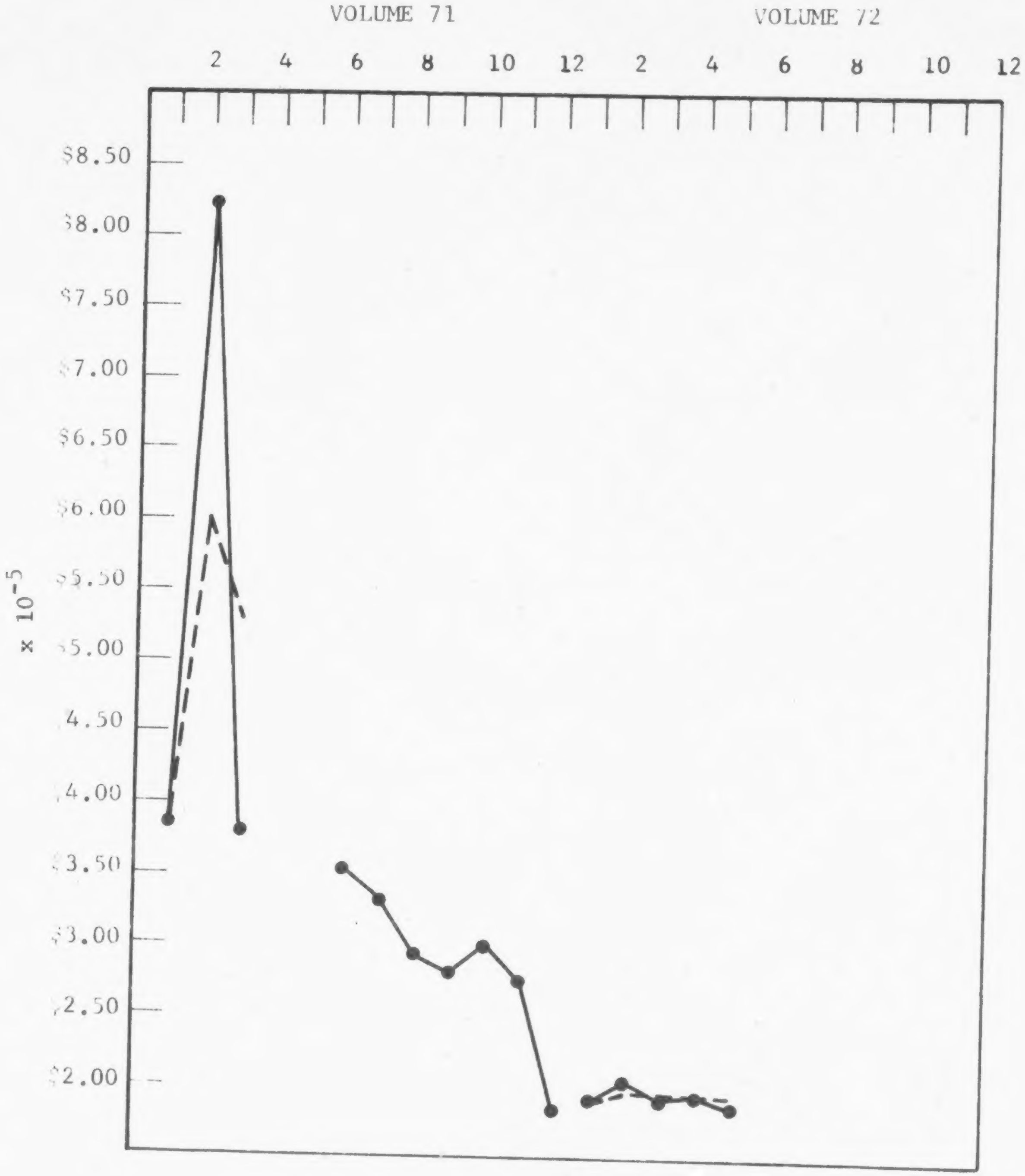


Figure 11-45

COST PER TERM PER CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71, 72

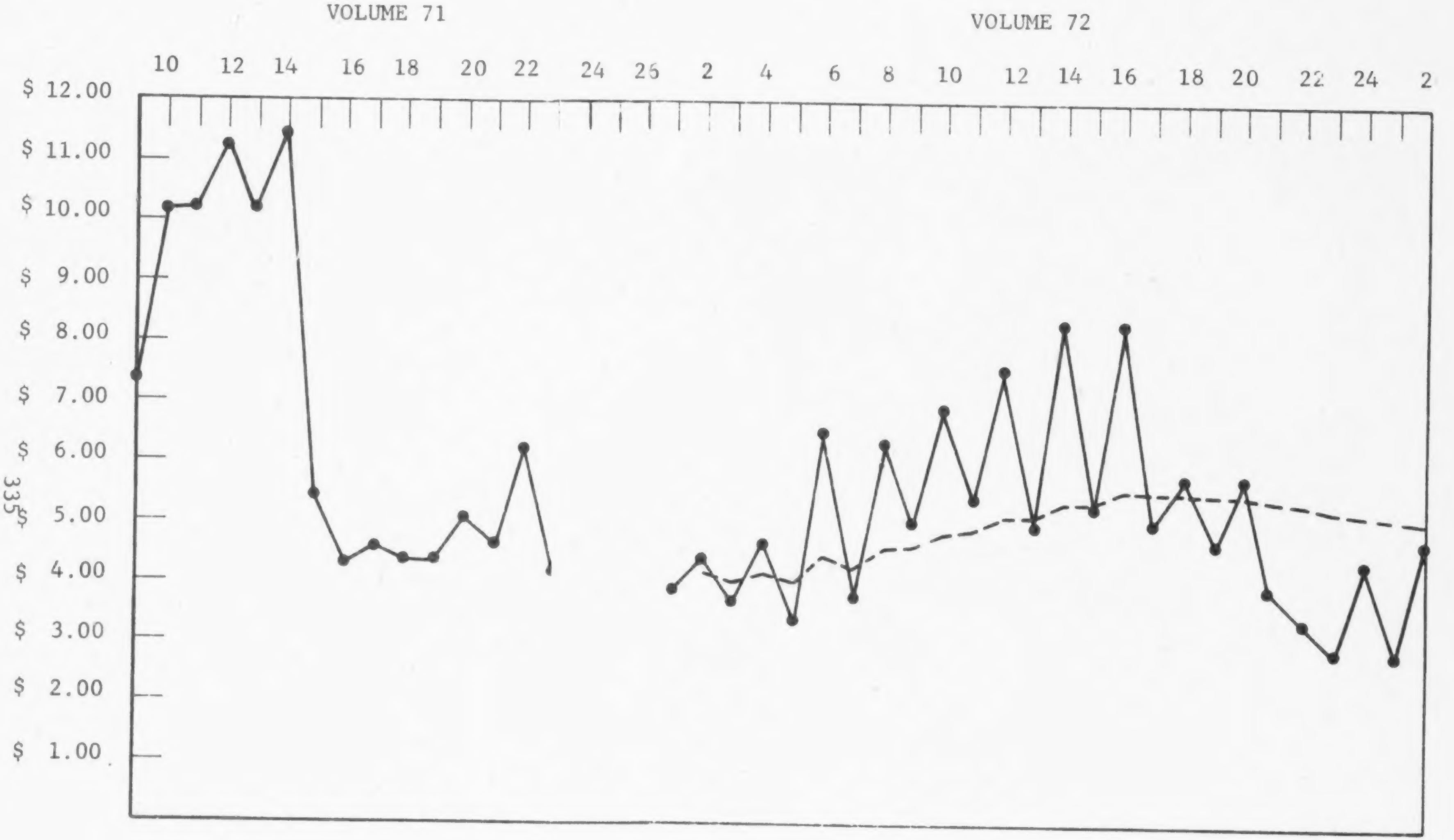


Figure 11-46

COST PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

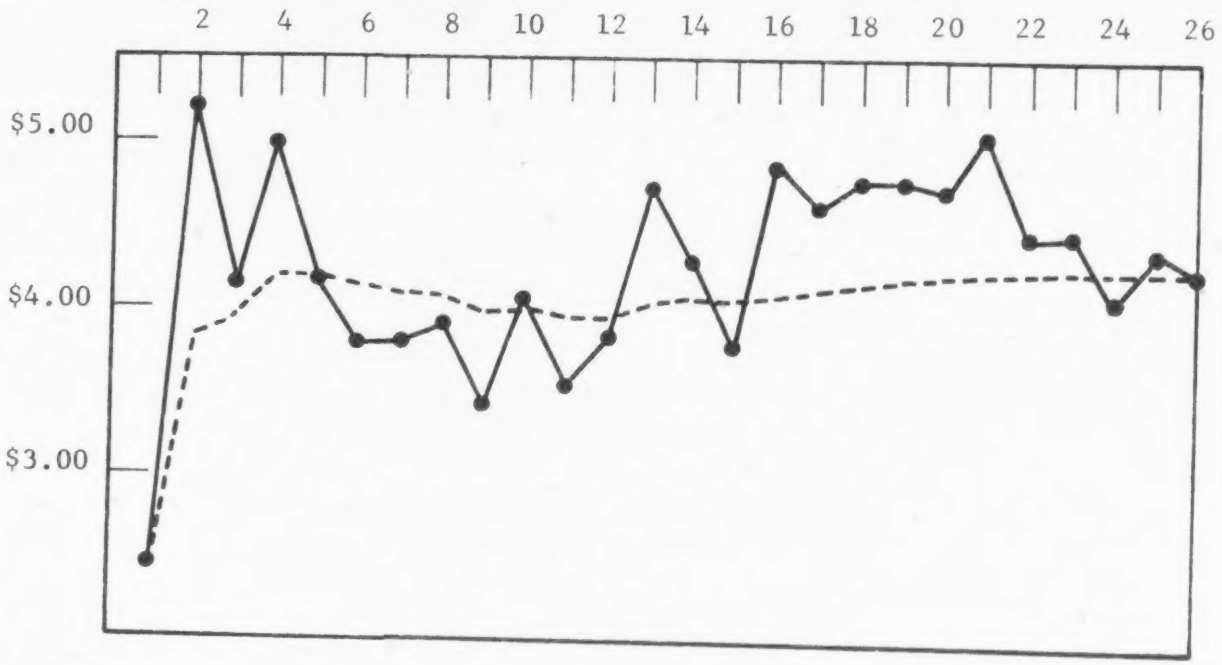


Figure 11-47

COST PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

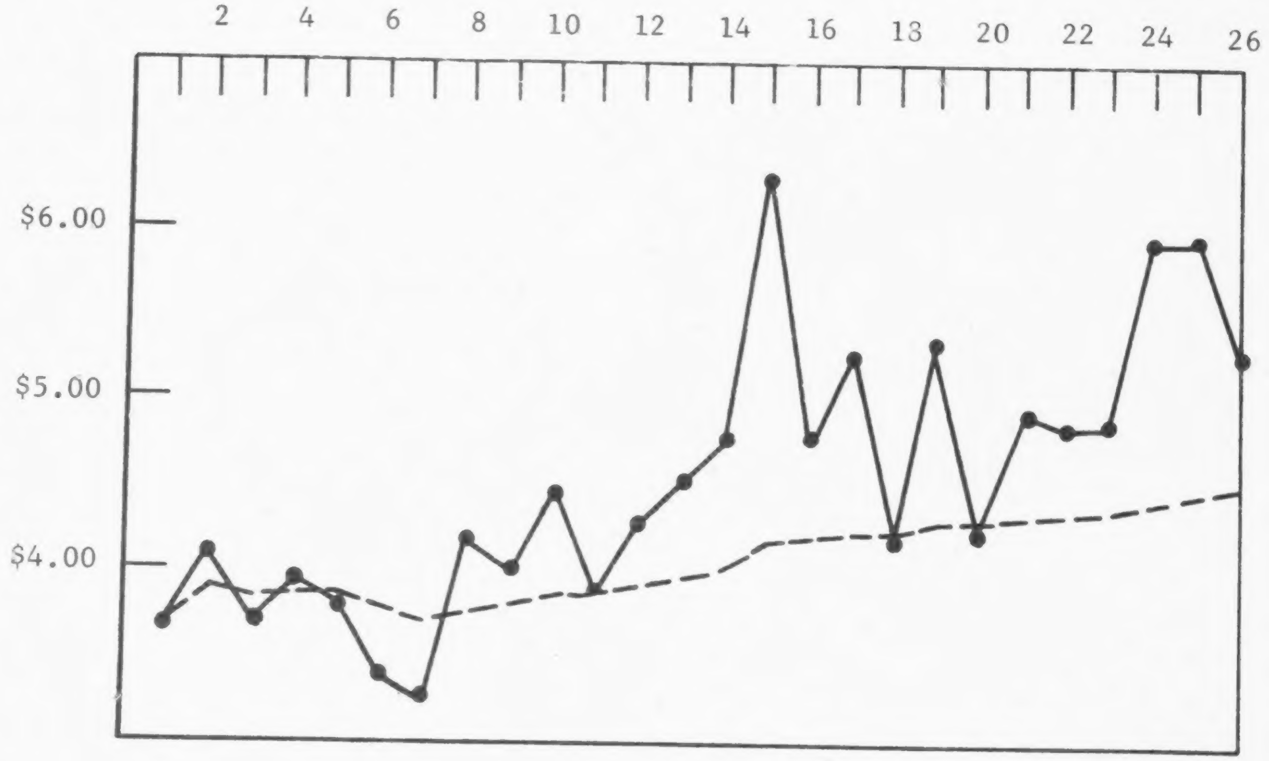


Figure 11-48

COST PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

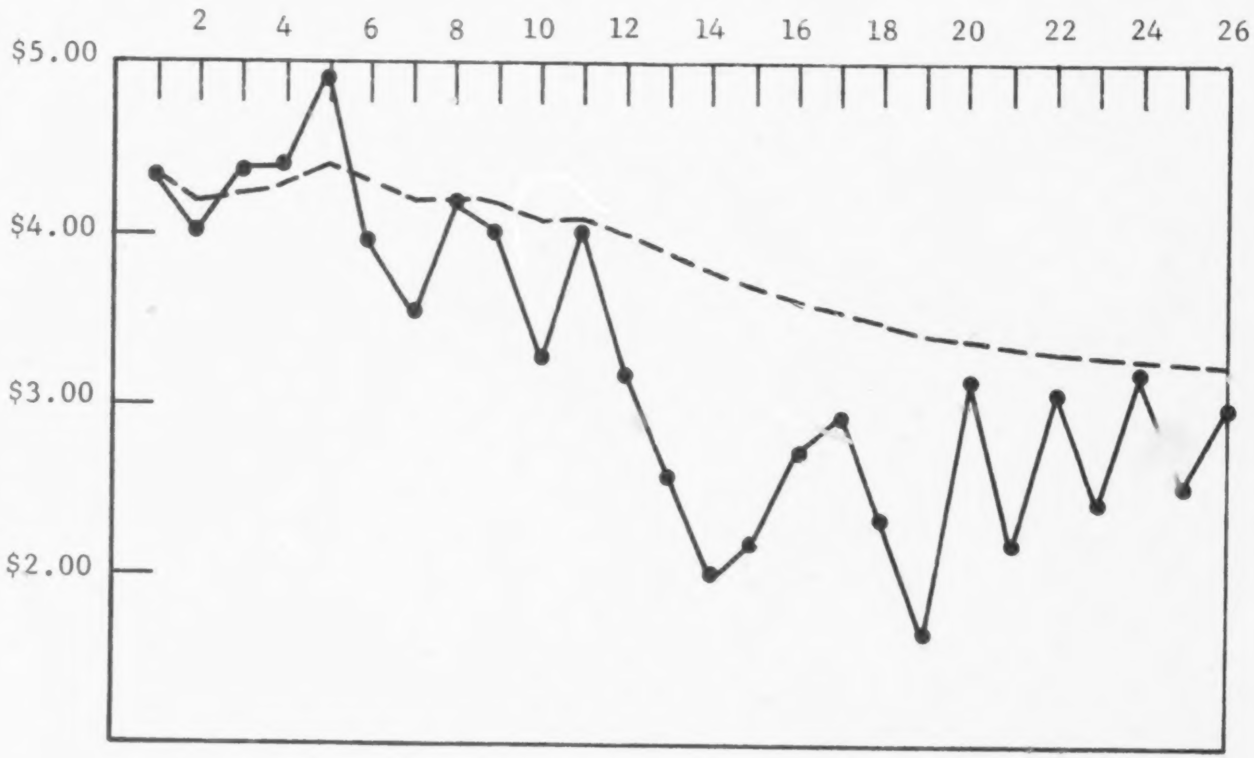


Figure 11-49

COST PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

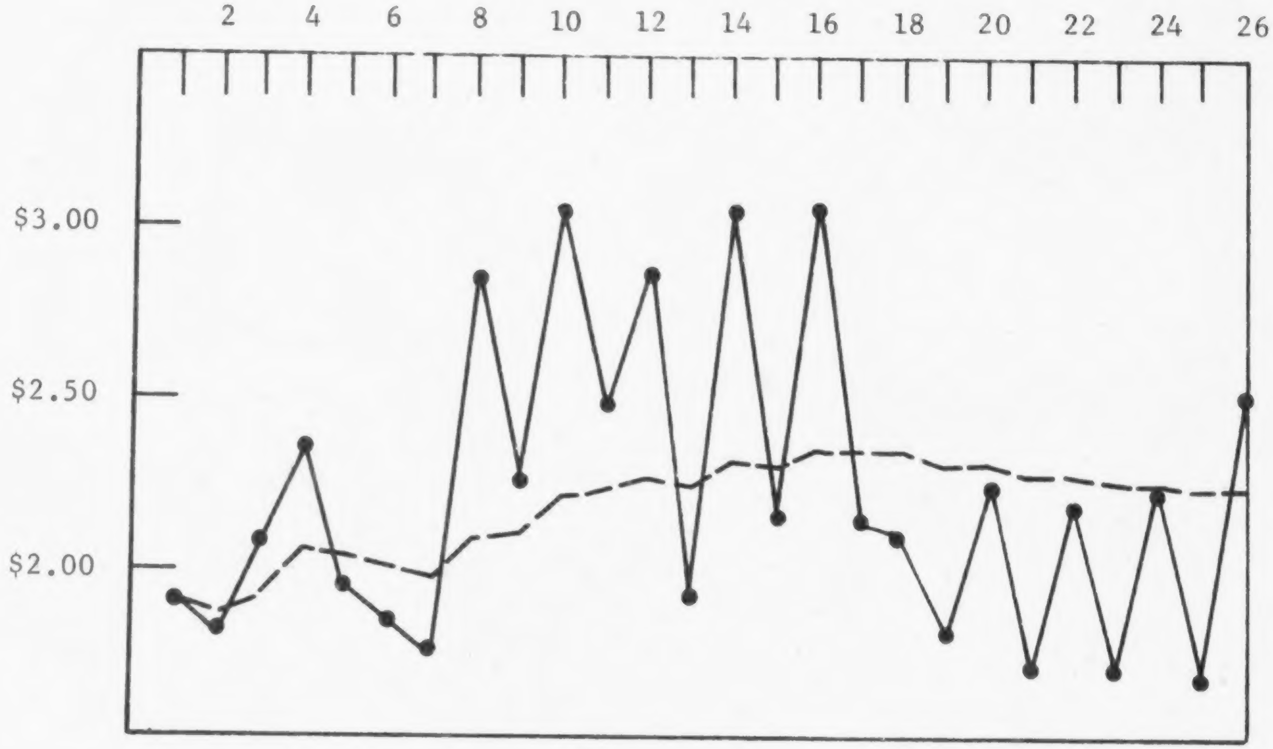


Figure 11-50

COST PER PROFILE VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

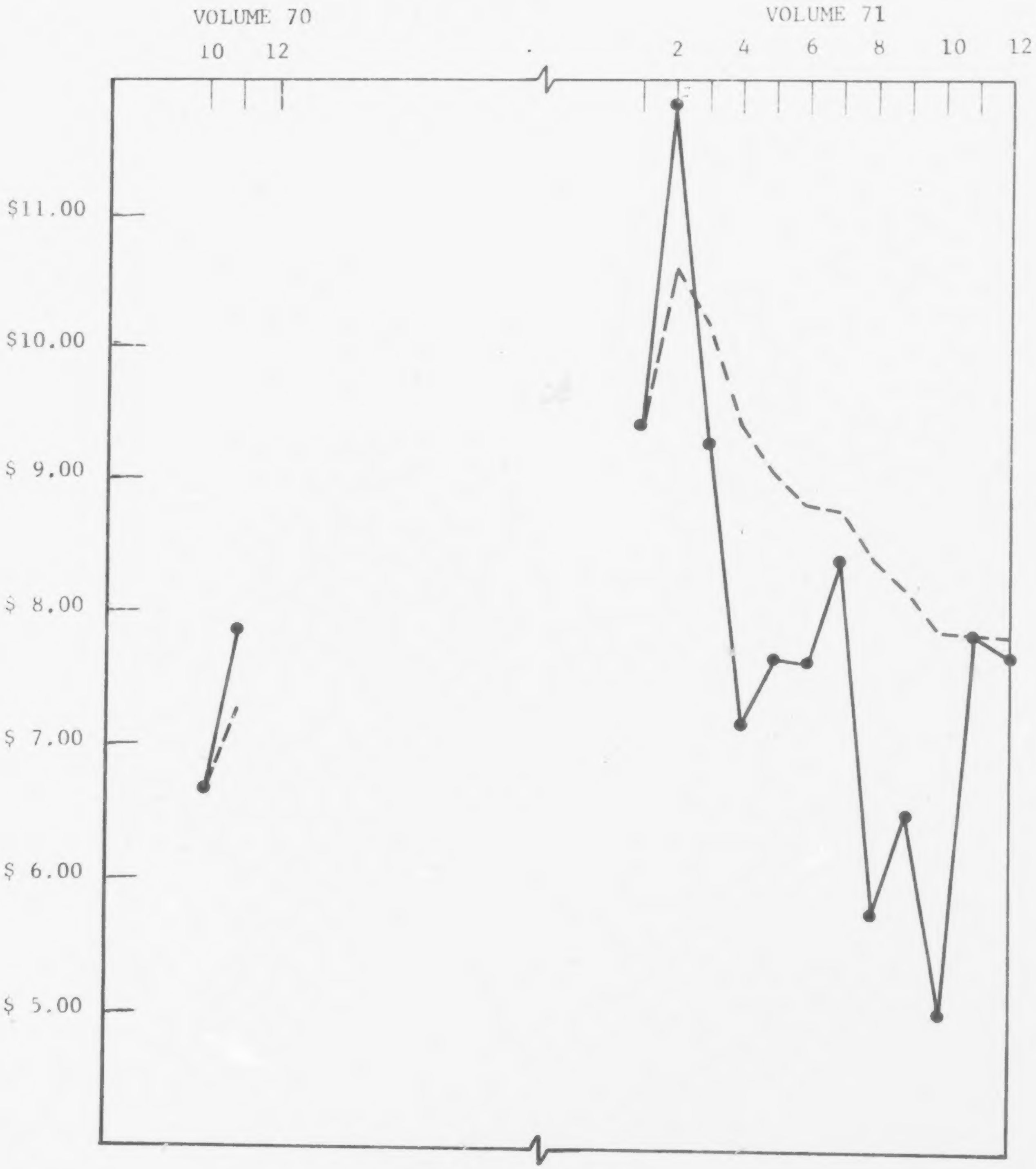


Figure 11-51

COST PER PROFILE VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

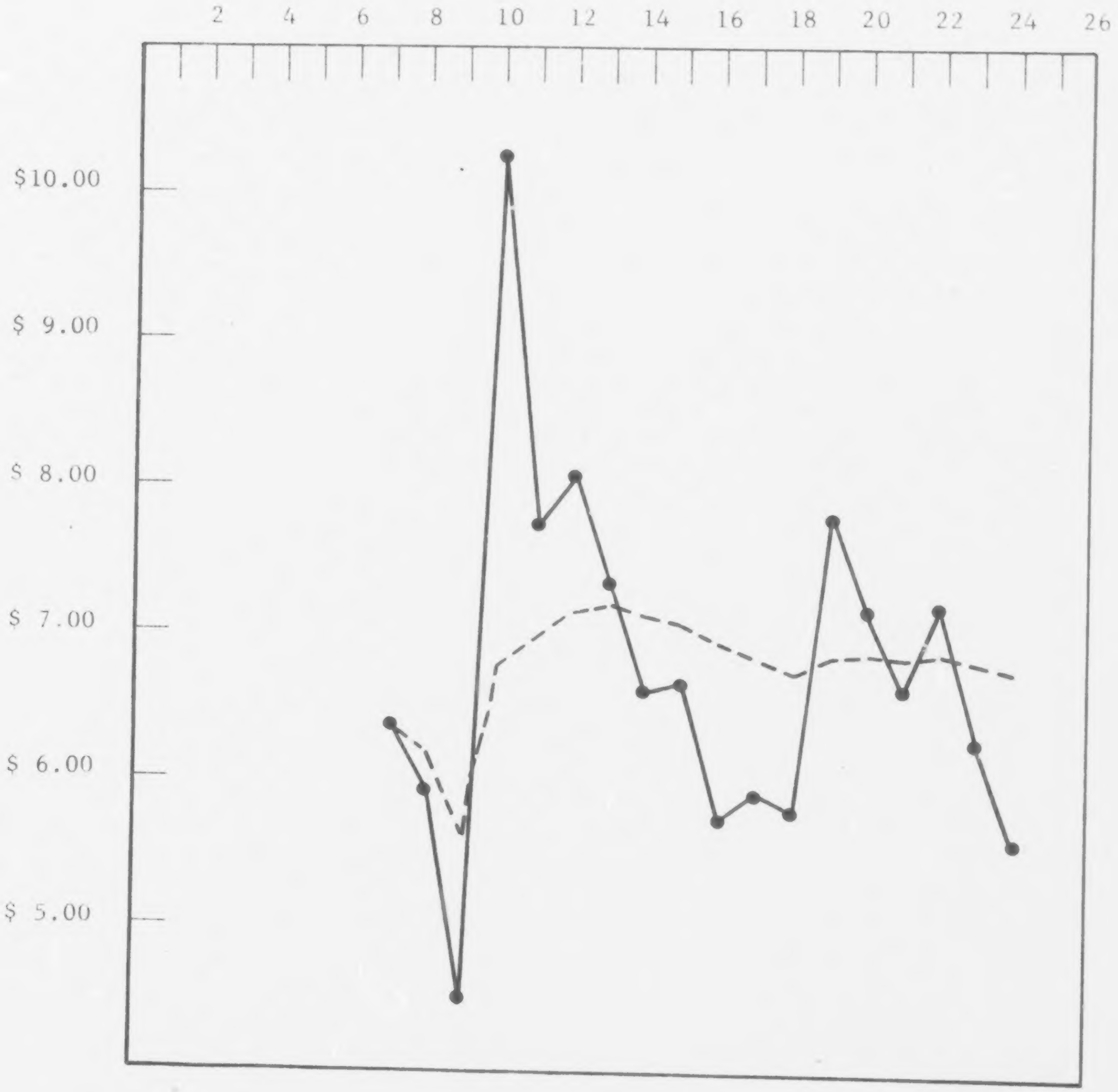


Figure 11-52

COST PER PROFILE VS. ISSUE

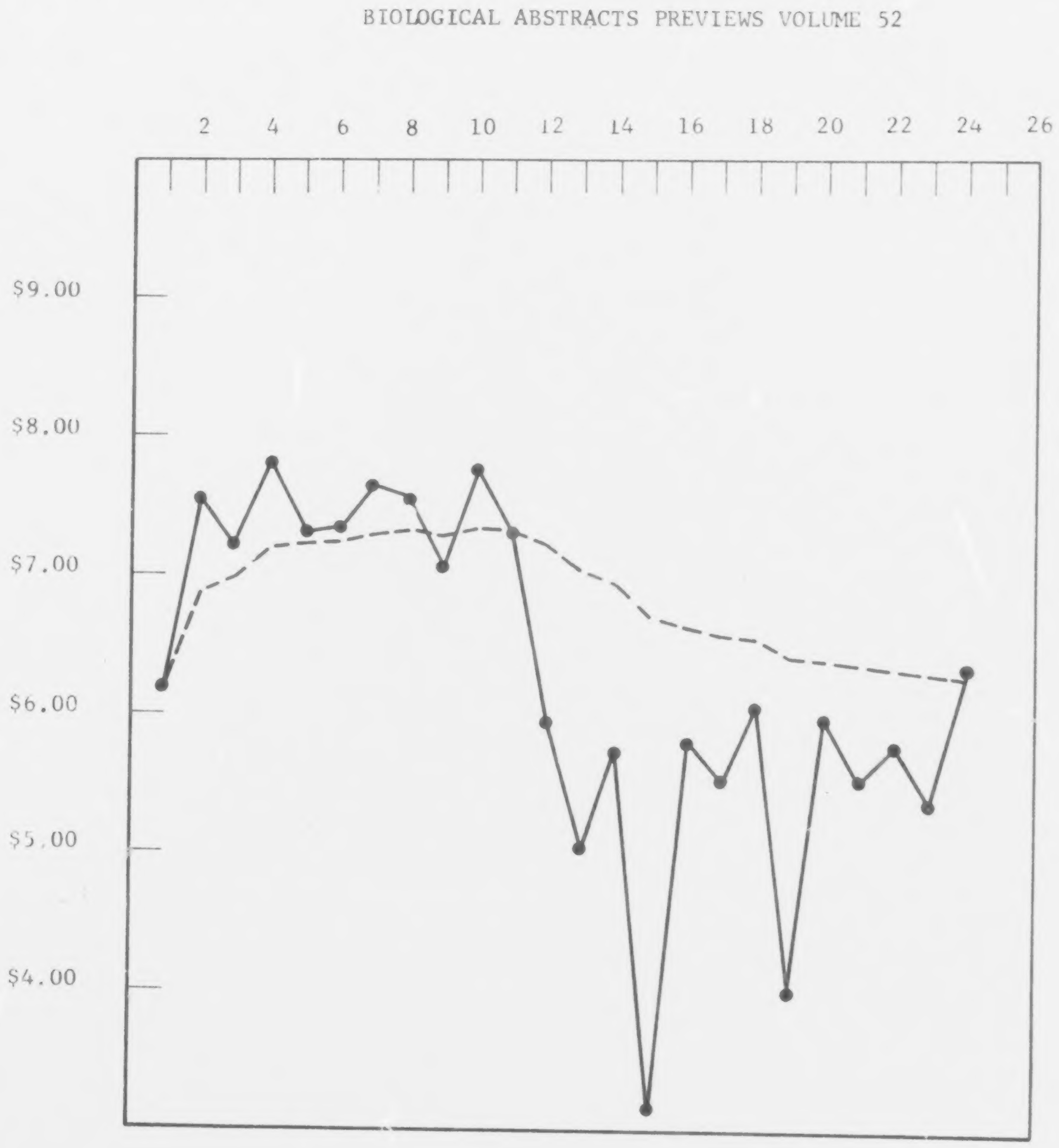


Figure 11-53

COST PER PROFILE VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUME 72

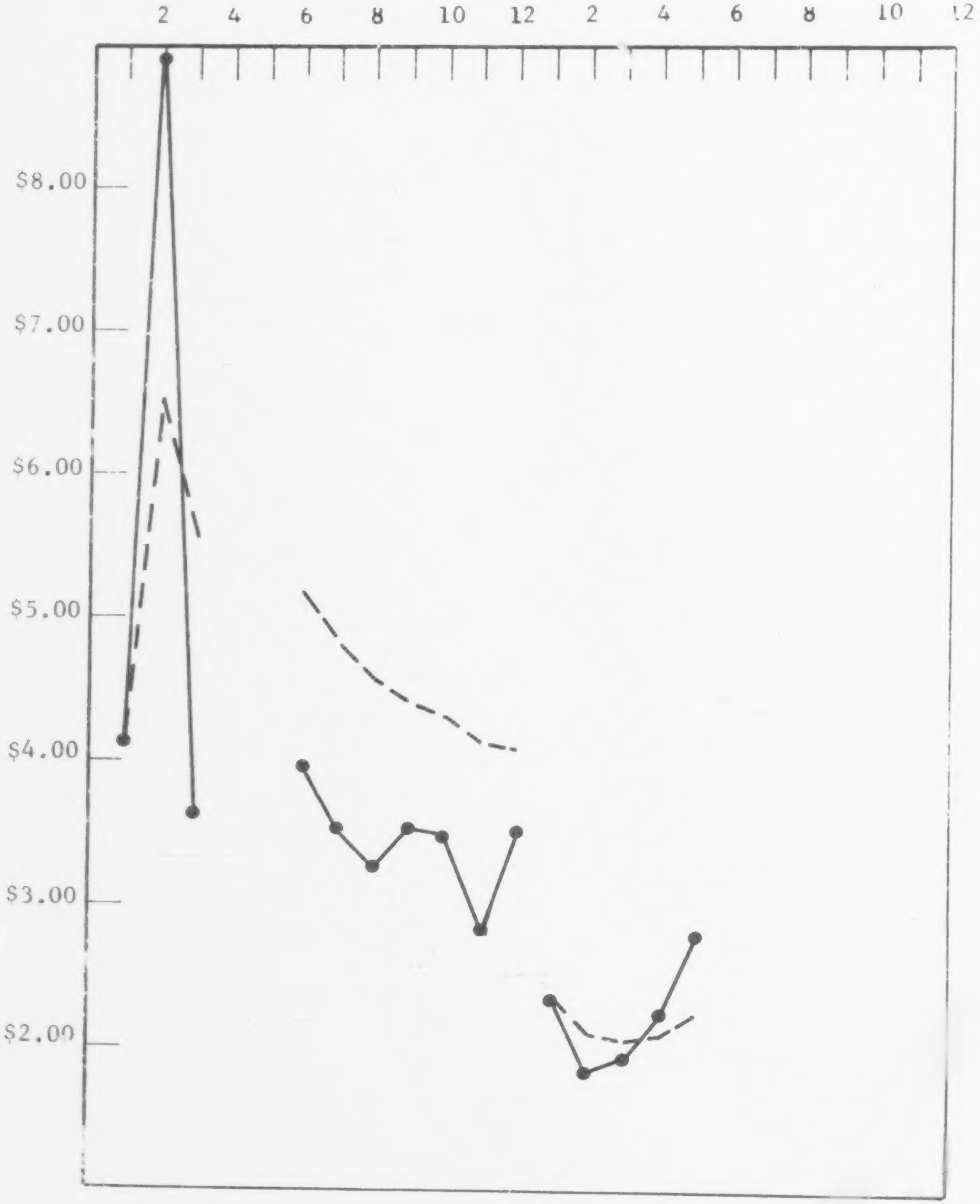


Figure 11-54

COST PER PROFILE VS. ISSUE

11.3 Hits (Output)

11.3.1 Hits--Profiles

CSC statistics generation programs produce data regarding the average numbers of hits per profile per issue of each data base--maximum, median and mean. The number of hits affects the royalties we pay to data base suppliers and hence our price structure. Some users cost us more in royalties because they generate more hits. With a print limit of 50 for the base subscription fee the average user is not constrained to try to cut down number of hits to avoid incurring added cost. The number of hits per profile per run ranges from 0 to 359. The average mean number of hits retrieved per profile per issue is 25 and the median is 16. This is dependent on data base size, hence a larger issue is likely to produce more hits per profile. This is true with the exception of maverick cases where inadvertantly a high frequency term is entered in an unrestricted manner thus generating an inordinate humber of hits for one profile.

The average number of hits per profile per issue for CA, BA, and EI are given in Figures 11-55 through 11-63, and normalized hits are presented in Figures 11-64 through 11-69. They are normalized to the average number of citations per issue for the volume in question.

While the mean number of hits per profile is 25 there are some profiles that get zero hits. Zero hit profiles can indicate several things:

(1) inappropriate data base,
(2) inappropriate issue of data base,
(3) overly specific terms,
(4) too tight logic, or
(5) desired output.

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71, 72

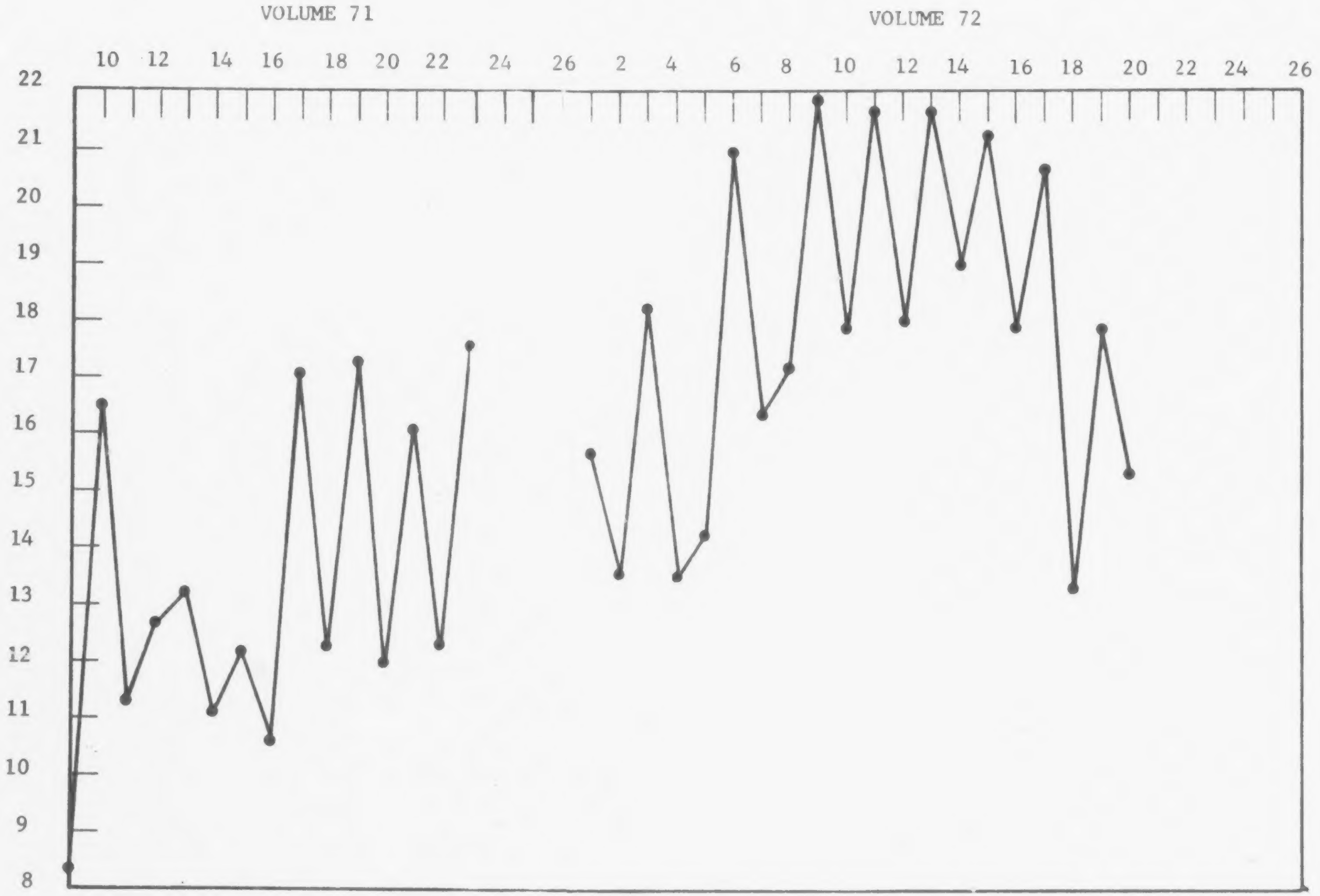


Figure 11-55

HITS PER PROFILE VS. ISSUE

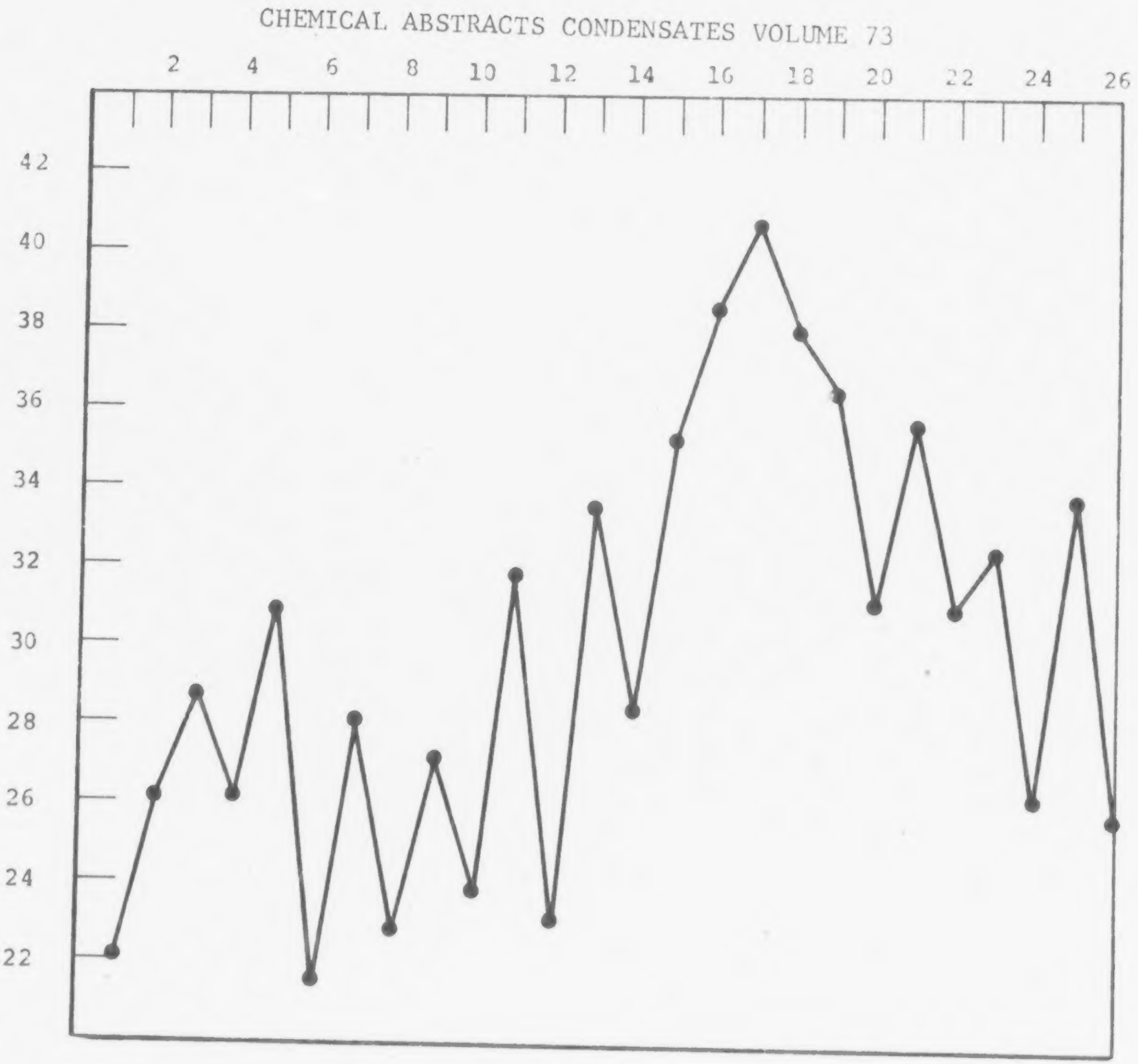


Figure 11-56

HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

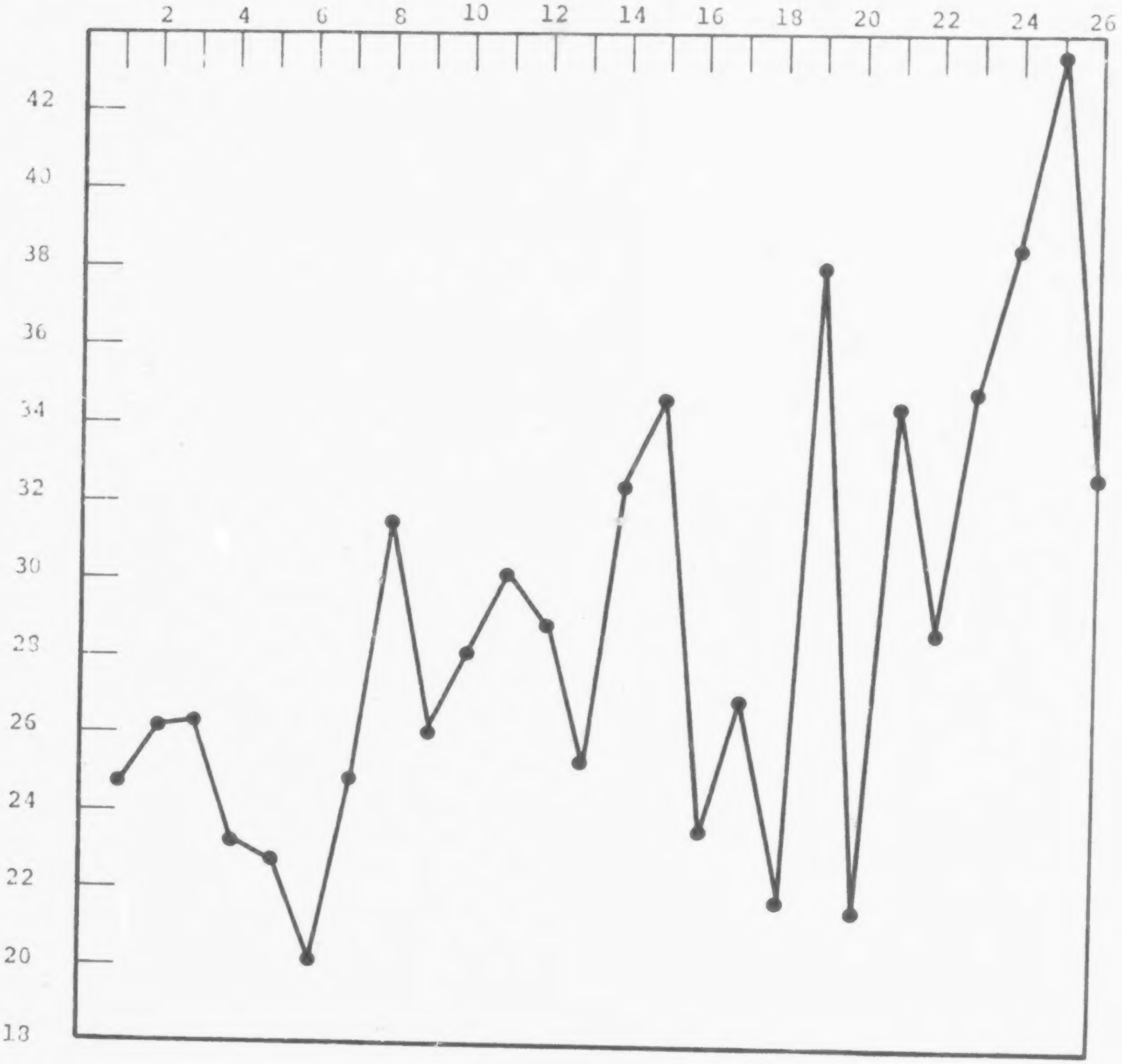


Figure 11-57

HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

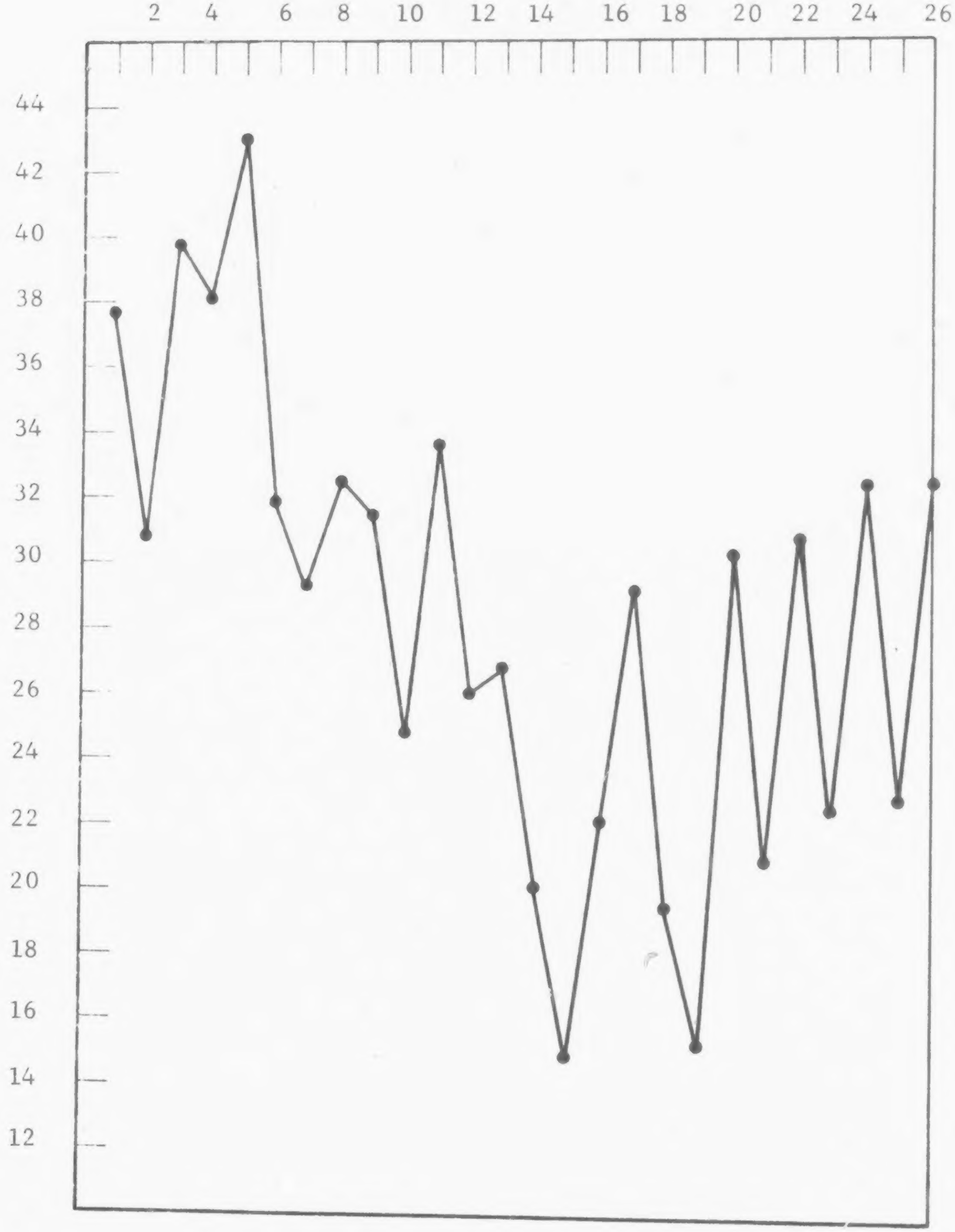


Figure 11-58

HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

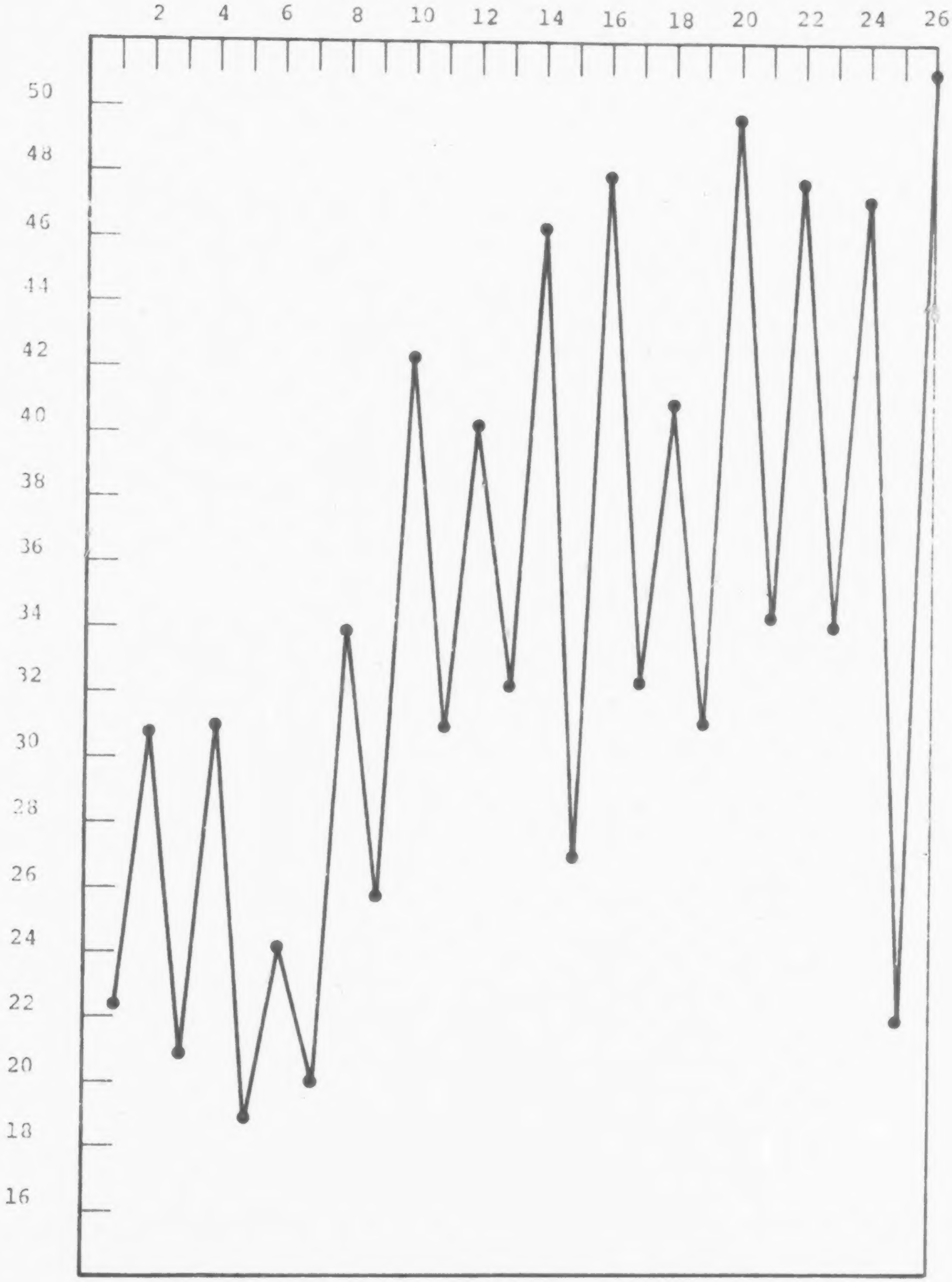


Figure 11-59

HITS PER PROFILE VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

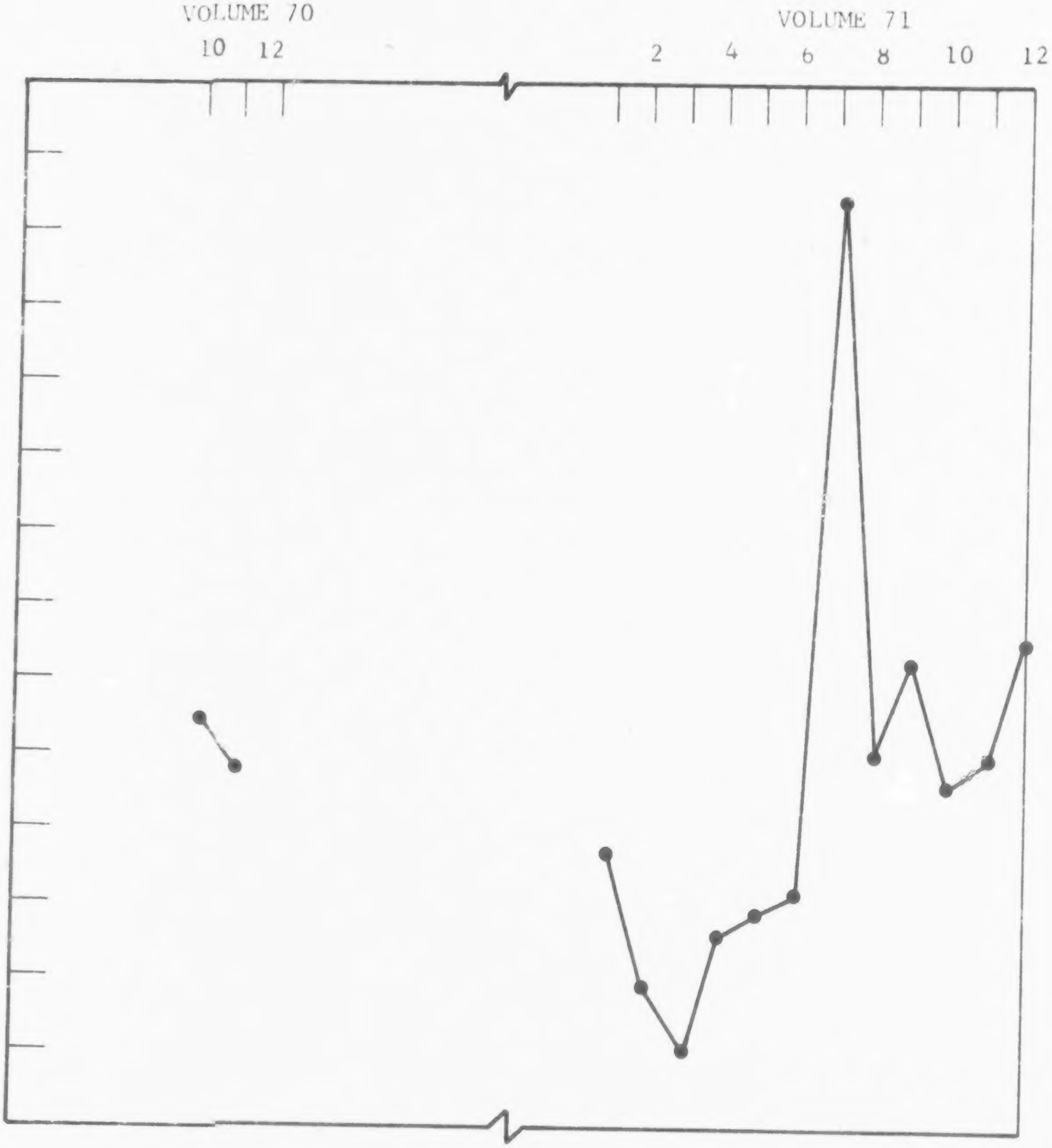


Figure 11-60

HITS PER PROFILE VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

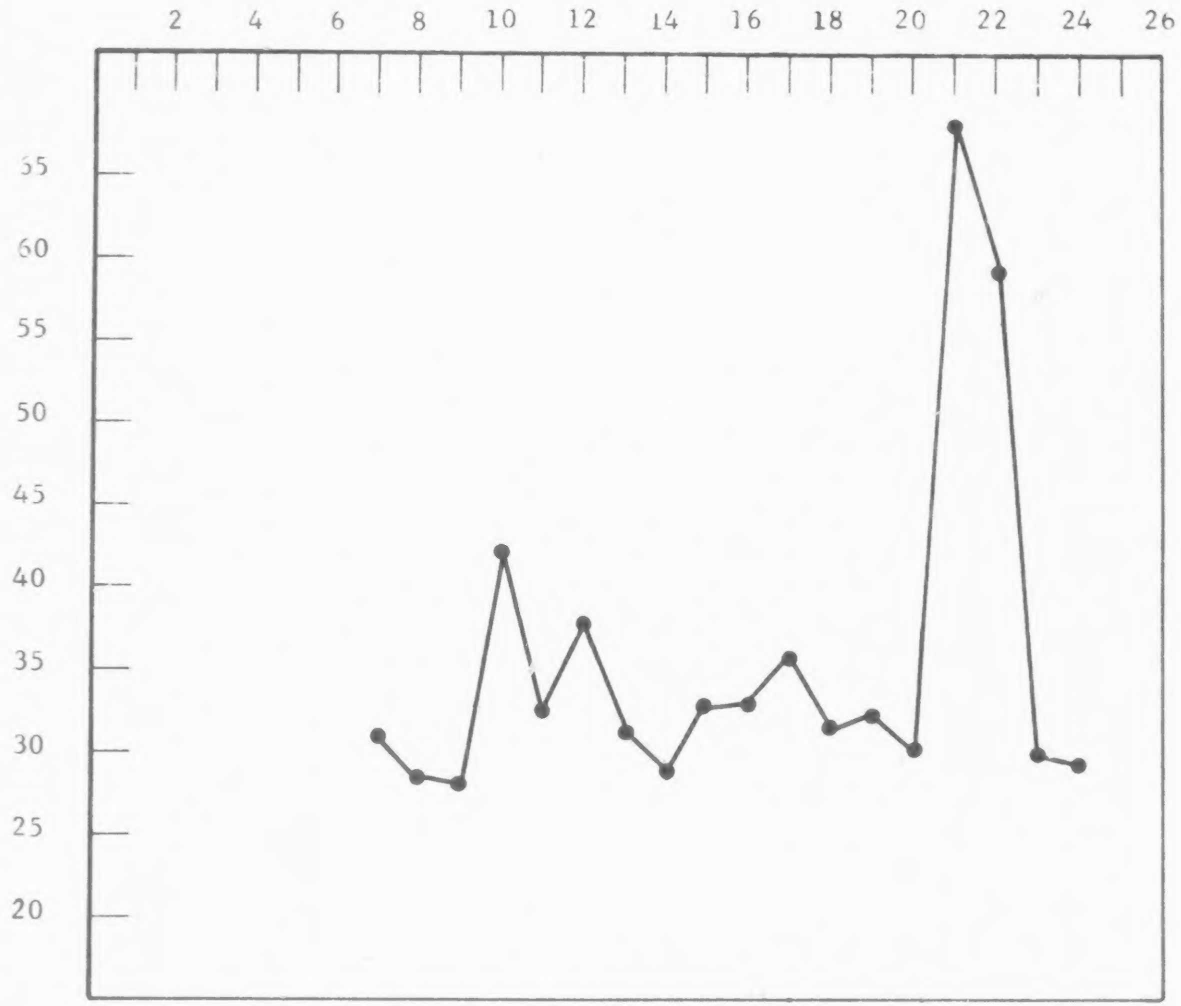


Figure 11-61

HITS PER PROFILE VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

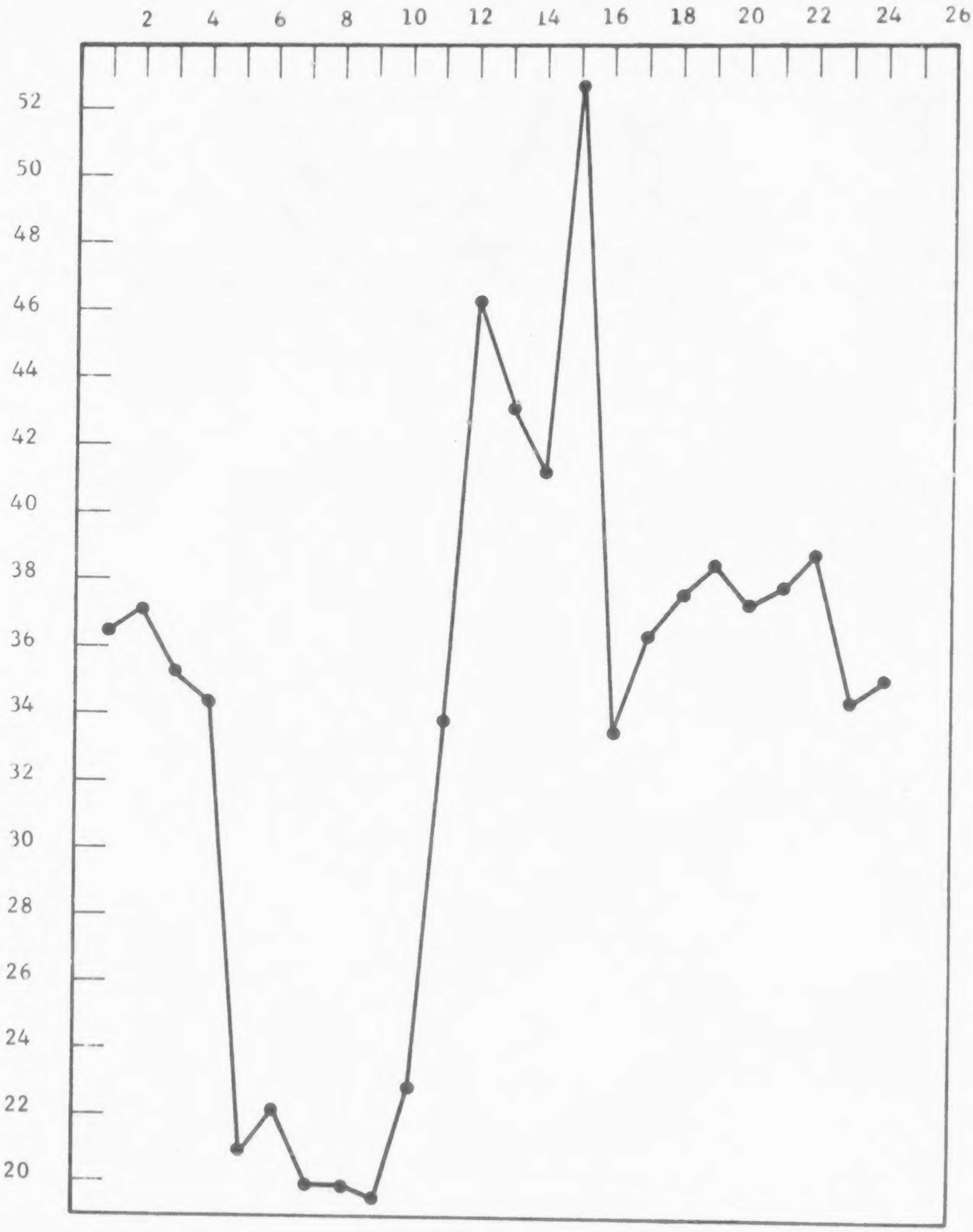


Figure 11-62

HITS PER PROFILE VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

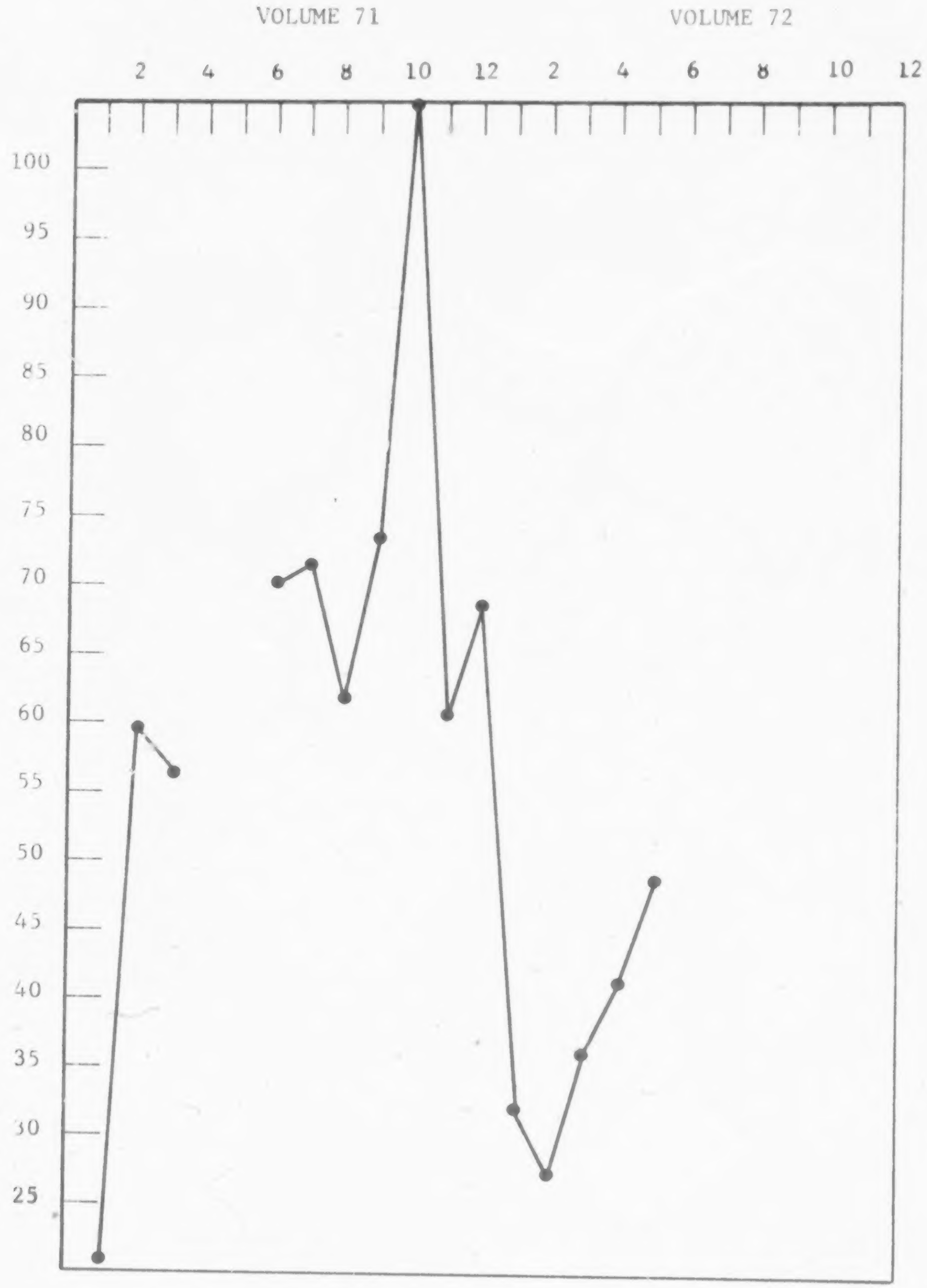


Figure 11-63

HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 71

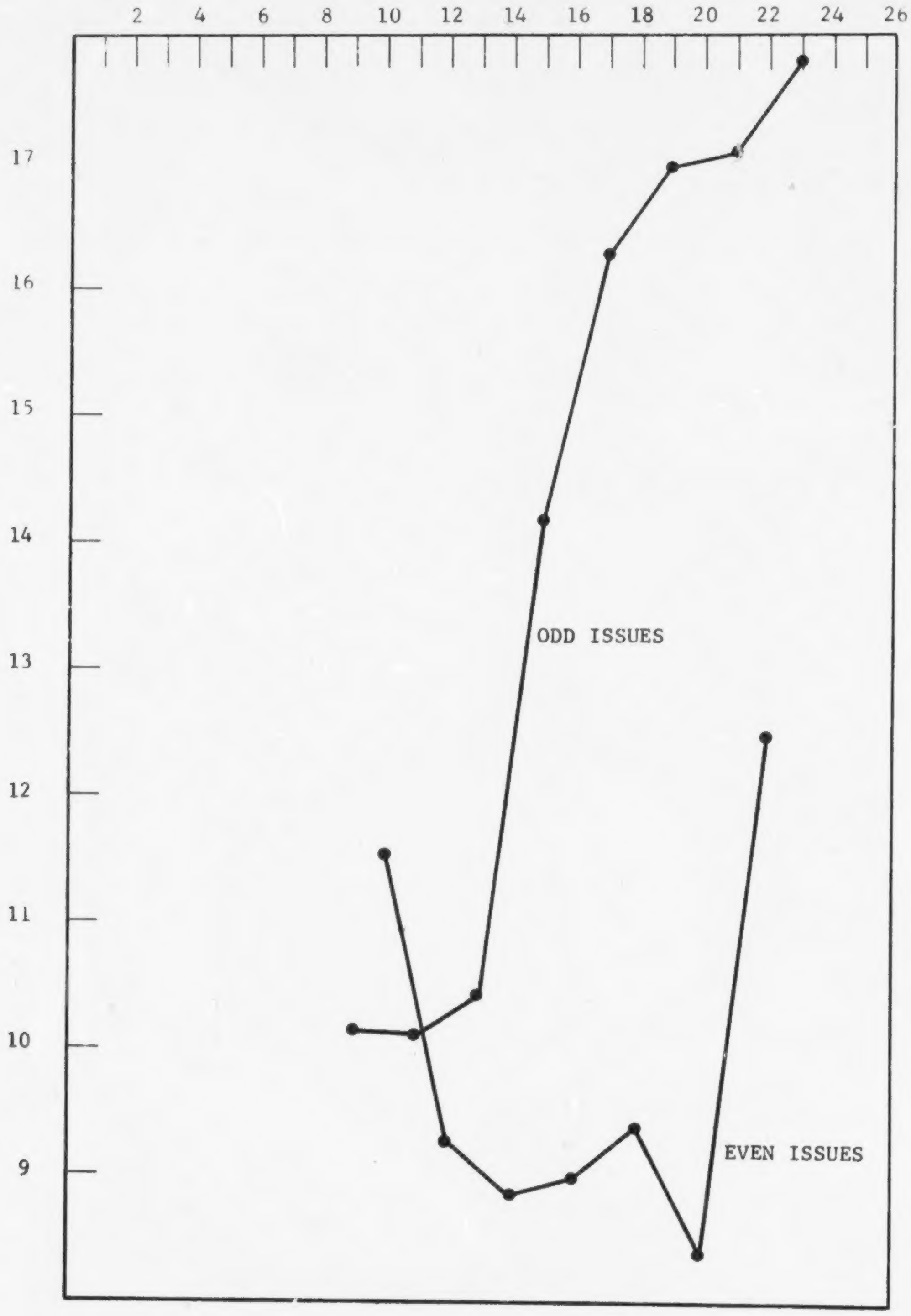


Figure 11-64
NORMALIZED HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

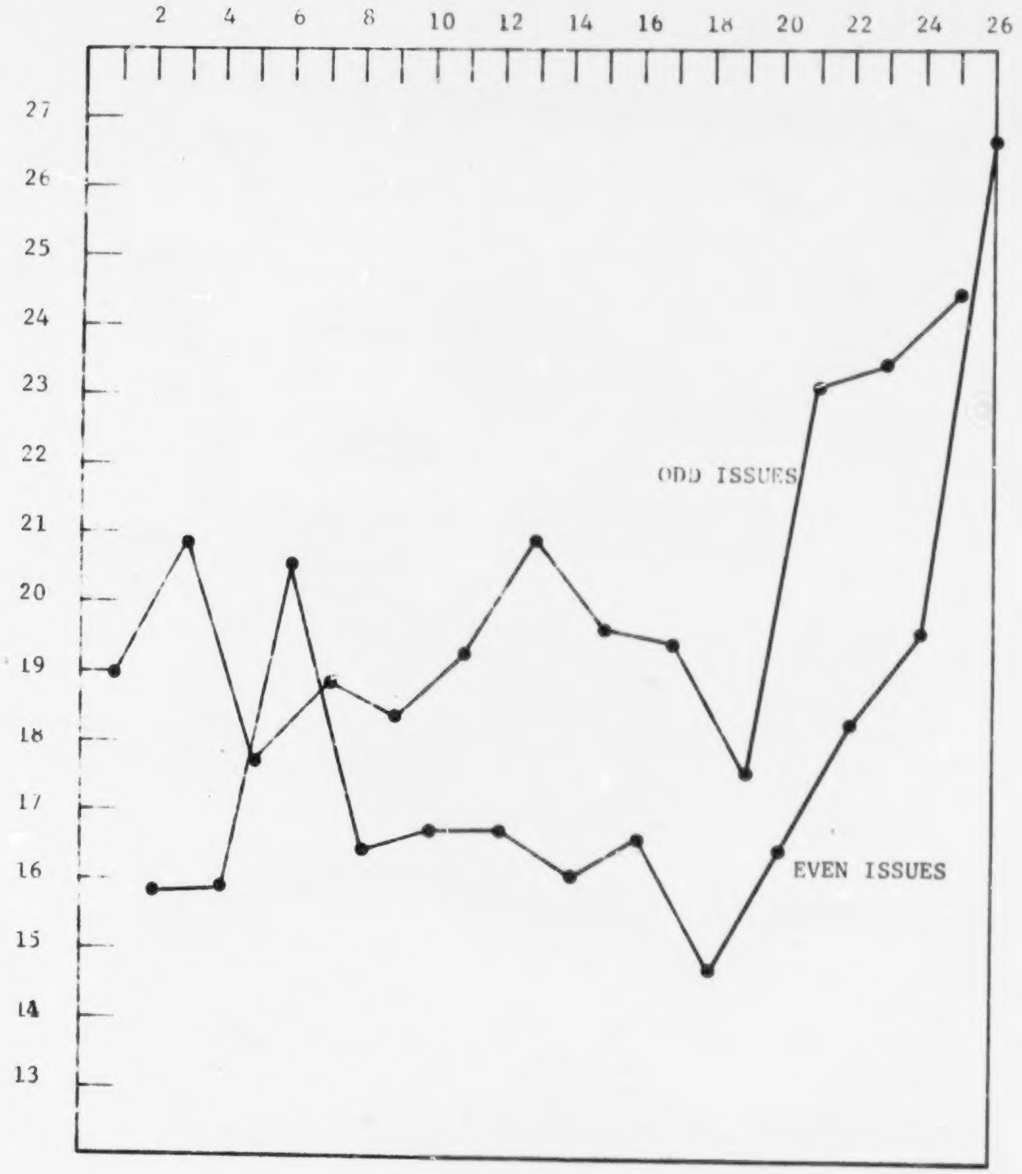


Figure 11-65

NORMALIZED HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

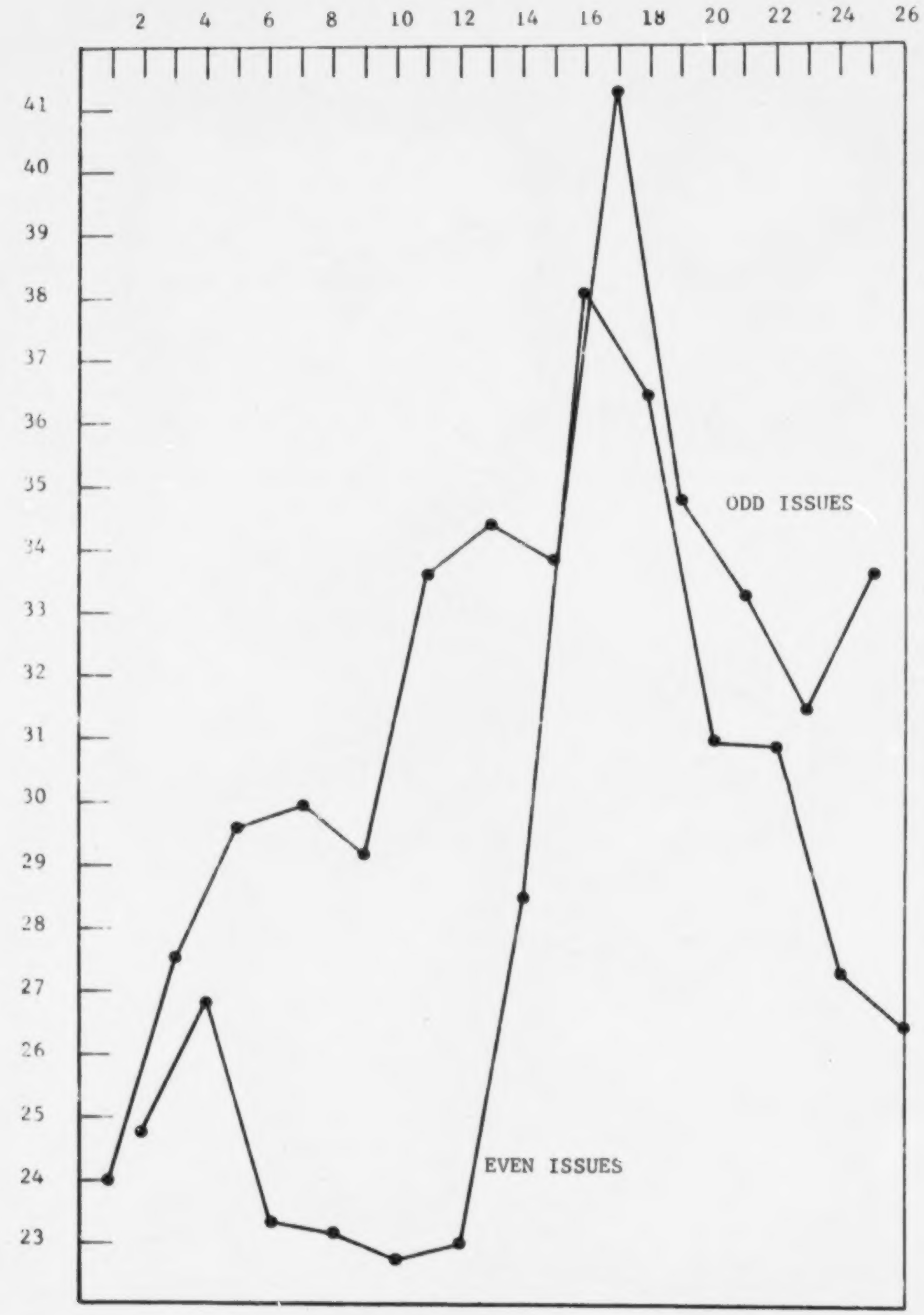


Figure 11-66

NORMALIZED HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

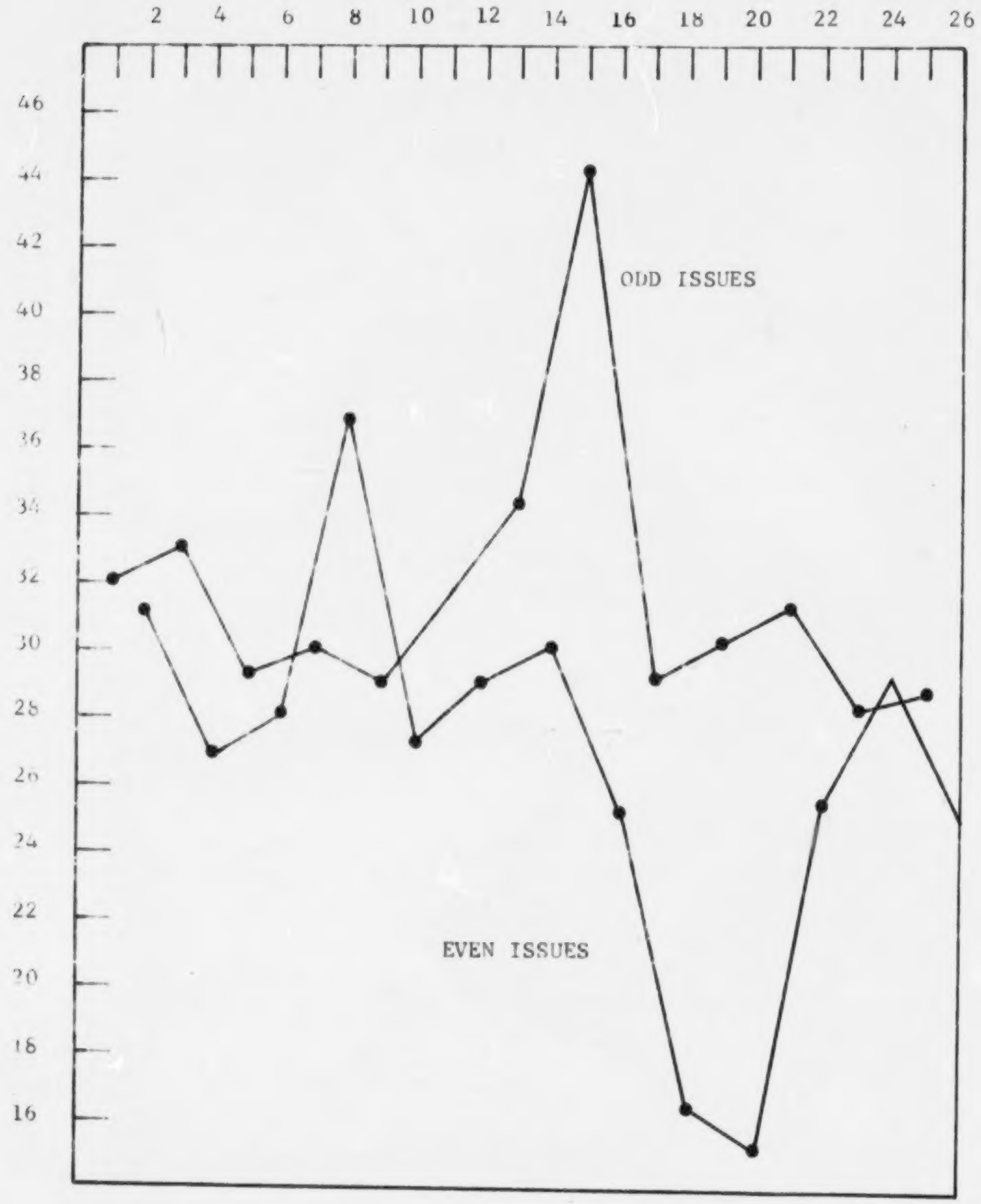


Figure 11-67

NORMALIZED HITS PER PROFILE VS. ISSUE

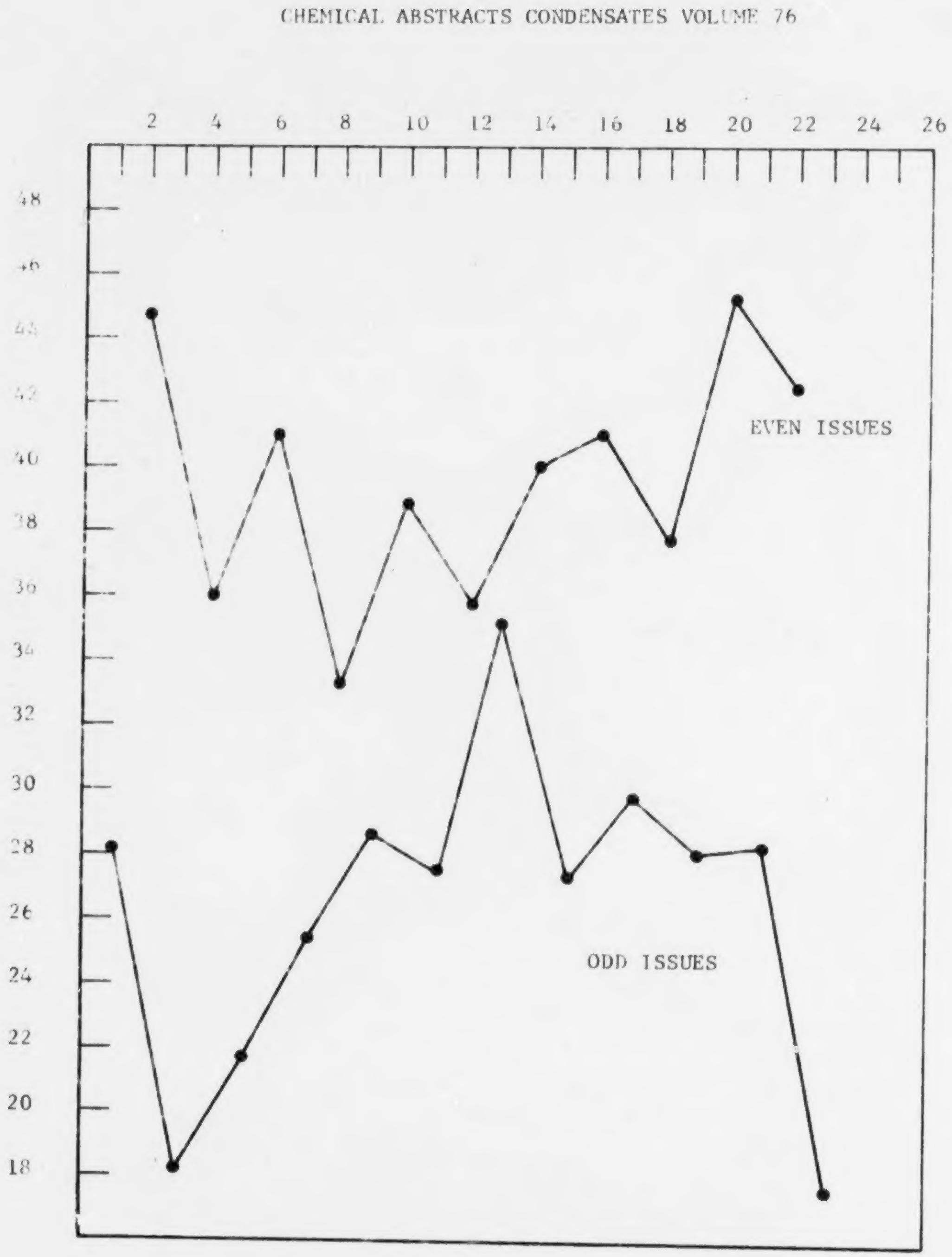


Figure 11-68

NORMALIZED HITS PER PROFILE VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

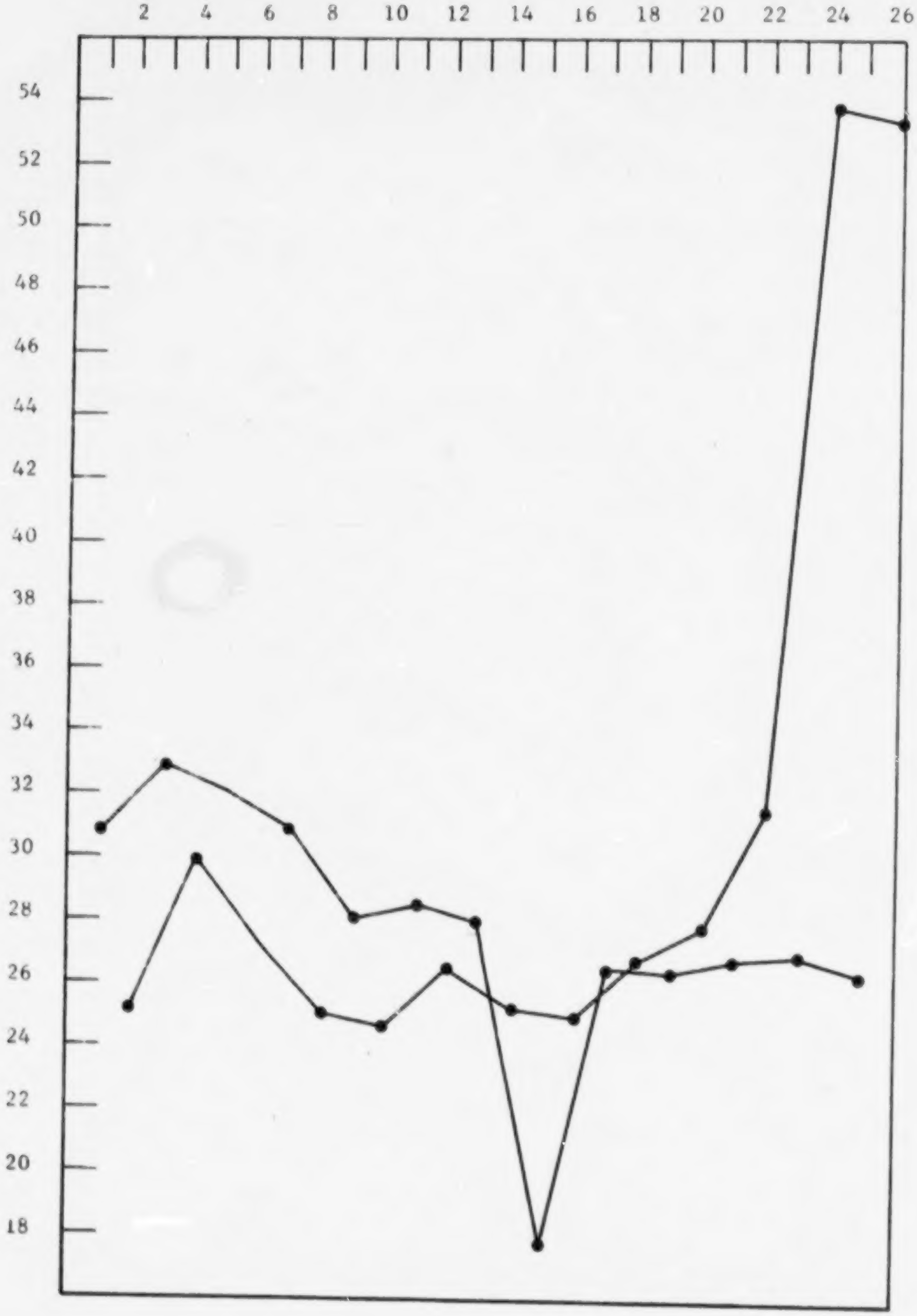


Figure 11-69

NORMALIZED HITS PER PROFILE VS. ISSUE

11.3.2 Hits--Retrieved Citations

The average number of hits per retrieved citation per issue tells how many of the profiles are retrieving the same citation and therefore indicates, to some extent, the homogeneity of the total user group.

Currently we are averaging 1.6 hits per retrieved citation per issue. If all hits were printed (and 95% are), the center would pay 1.6 times the royalty fee for a given hit. Appriximately 60% of the citations in the data base are found as hits for one or more profiles. This is true of CA, BA, and EI. Naturally the number of profiles must reach approximately 100 for this to be true. After that point the percentage does not seem to increase, though seemingly with an extremely large number of heterogeneous profiles the percentage would probably increase asymptotically to 99+. Unfortunately we have not had the opportunity of checking this out.

The number of citations that are hits is probably a function of the heterogeneity of the profile group. Or, it may be related to the fact that some citations have titles with no definitive terms and are also poorly indexed.

Hits per retrieved citation per issue for CA, BA, and EI are given in Figures 11-70 through 11-78. Normalized hits per issue are given in Figures 11-79 through 11-84 and CSC machine cost per hit per issue is given in Figures 11-85 through 11-93.

11.3.3 Printed Hits

Not all citations that are hits for a profile are necessarily printed. A user may specify a print limit. Though most hits are printed some are not. The mean number of prints per profile per issue is 23.5 and the median is 15.

The number of prints affects center cost somewhat but printing cost is minimal (2% of total run) as we print offline at significantly lower rates. We print approximately 150 K lines/week. While printing cost is low, postage for shipping large number's of cards at 1st class rates and the purchase of the card stock is a real cost;

e.g.,	card	8.0 mils
	postage	4.0 mils
	print	1.1 mils
		13.1 mils or 31.2¢ per profile.

The number of hits and prints generated by each user organization is a statistic that is automatically generated. The corporate distribution of hits and prints is the same as profile hit distribution. It shows which companies are generating high numbers of hits (i.e., costing more) and it is tabulated by profile within company. This is an indicator for the center or user-company profile-coordinator as to which profiles are generating how many hits.

CSC uses these data in estimating profile subscriber fees for the next year; e.g., with a print limit of 50 for the base fee and added cost thereafter, a user who gets a large number of hits can predict the number of dollars he will need for the next year.

CHEMICAL ABSTRACTS CONDENSATES VOLUME 72

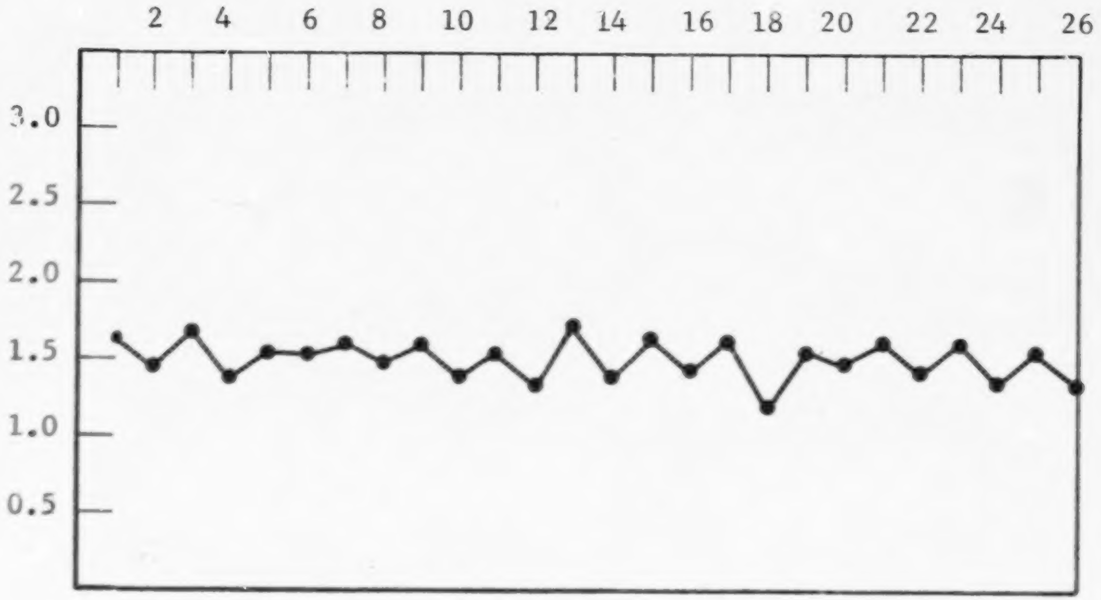


Figure 11-70

HITS PER RETRIEVED CITATION VS. ISSUE

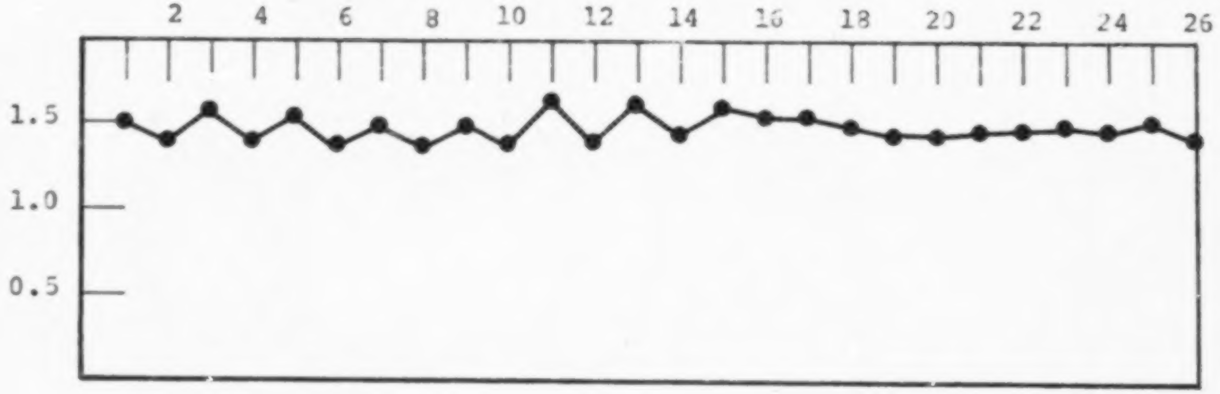

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

Figure 11-71

HITS PER RETRIEVED CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

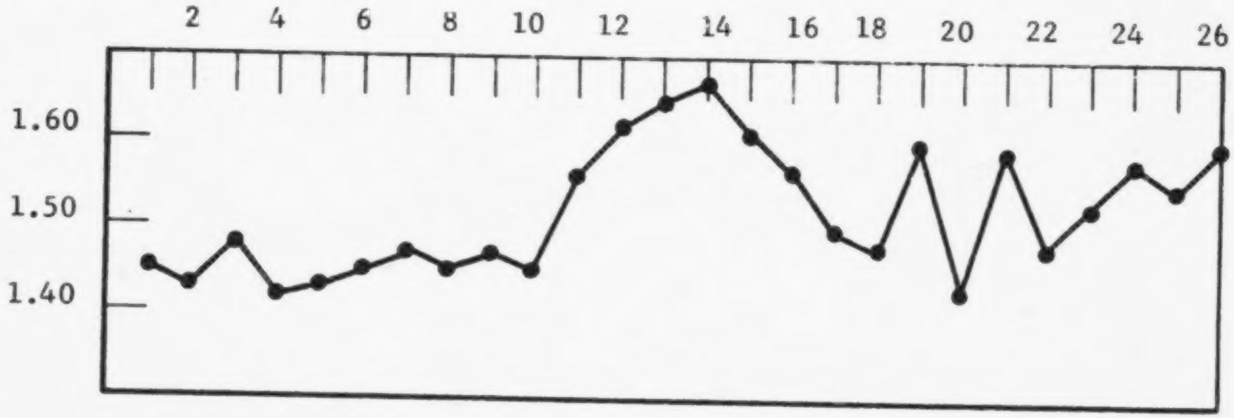


Figure 11-72

HITS PER RETRIEVED CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

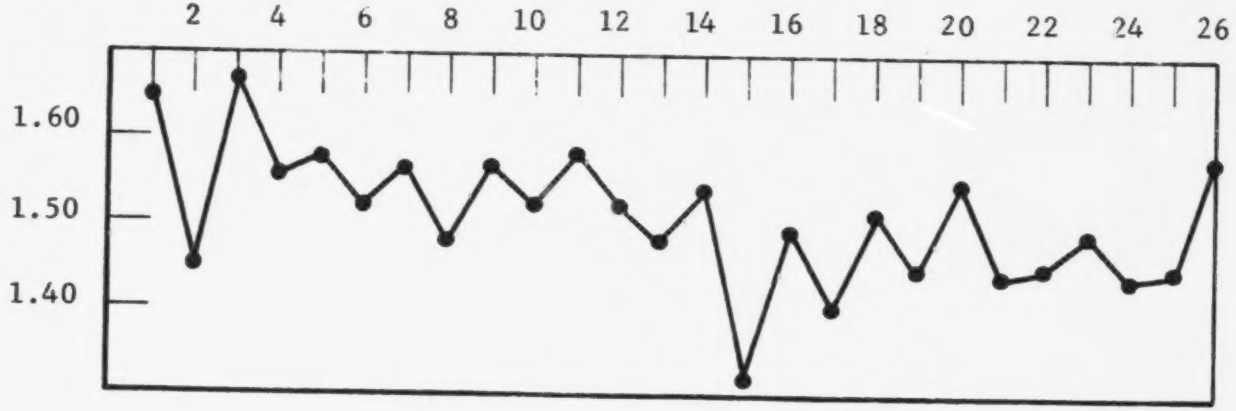


Figure 11-73

HITS PER RETRIEVED CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

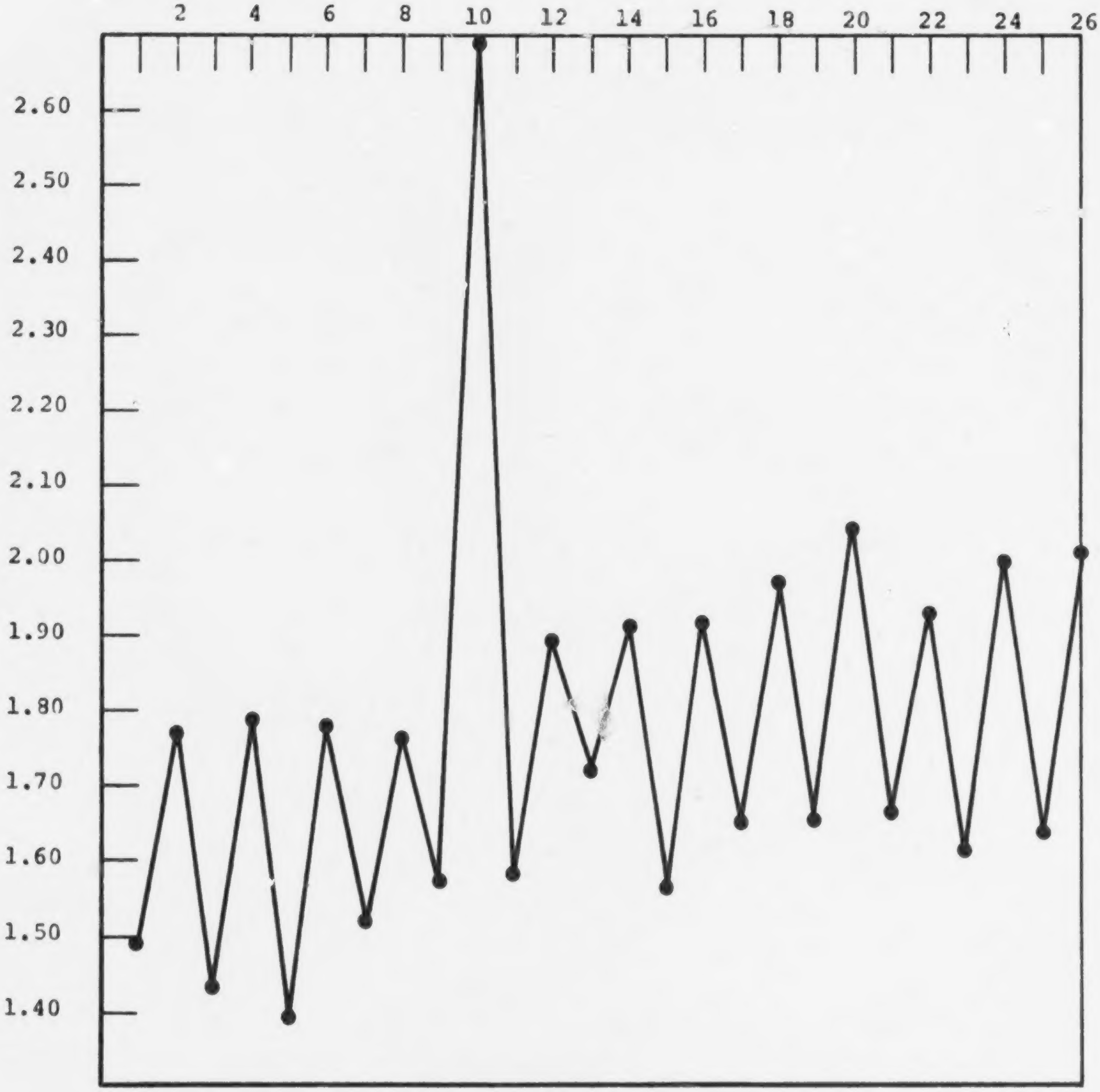


Figure 11-74

AVERAGE HITS PER RETRIEVED CITATION VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

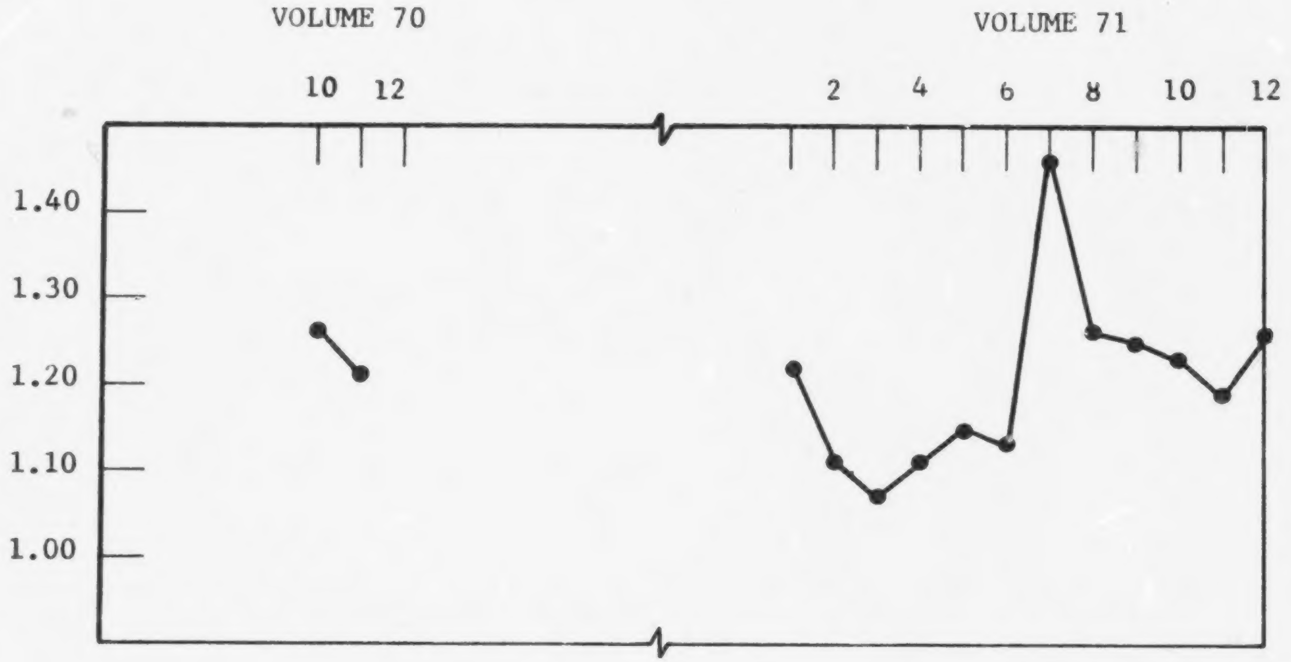


Figure 11-75

HITS PER RETRIEVED CITATION VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

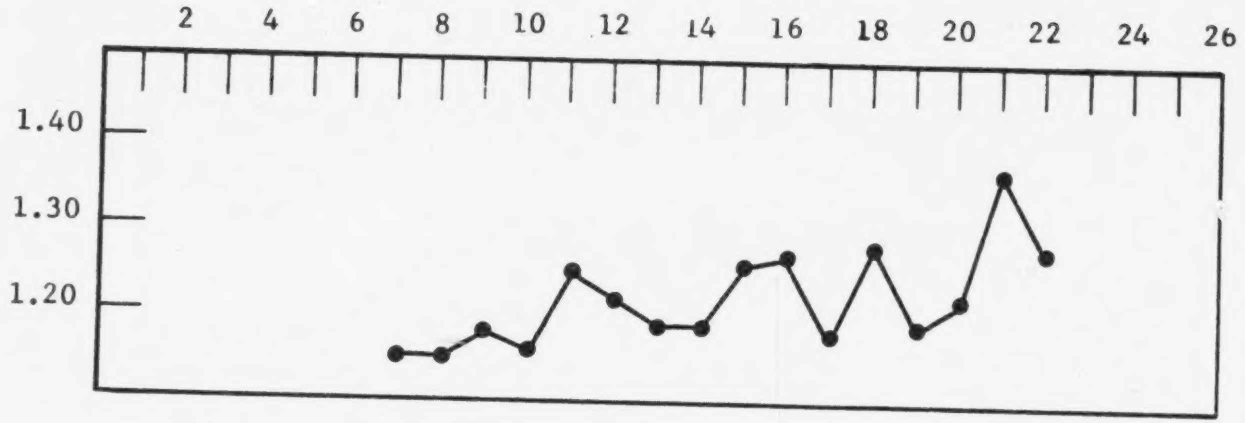


Figure 11-76

HITS PER RETRIEVED CITATION VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 52

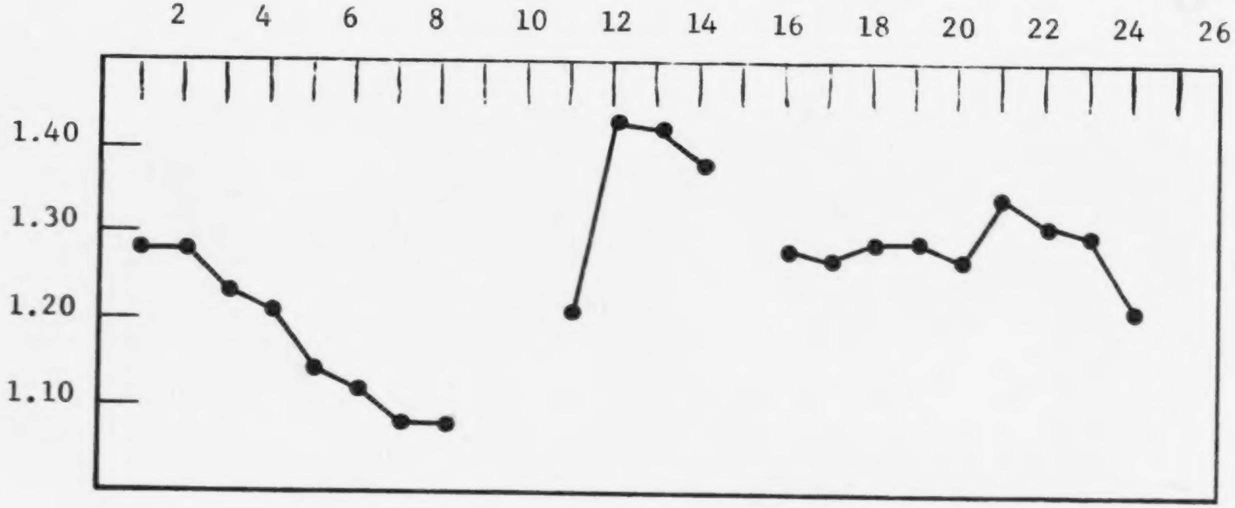


Figure 11-77

HITS PER RETRIEVED CITATION VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

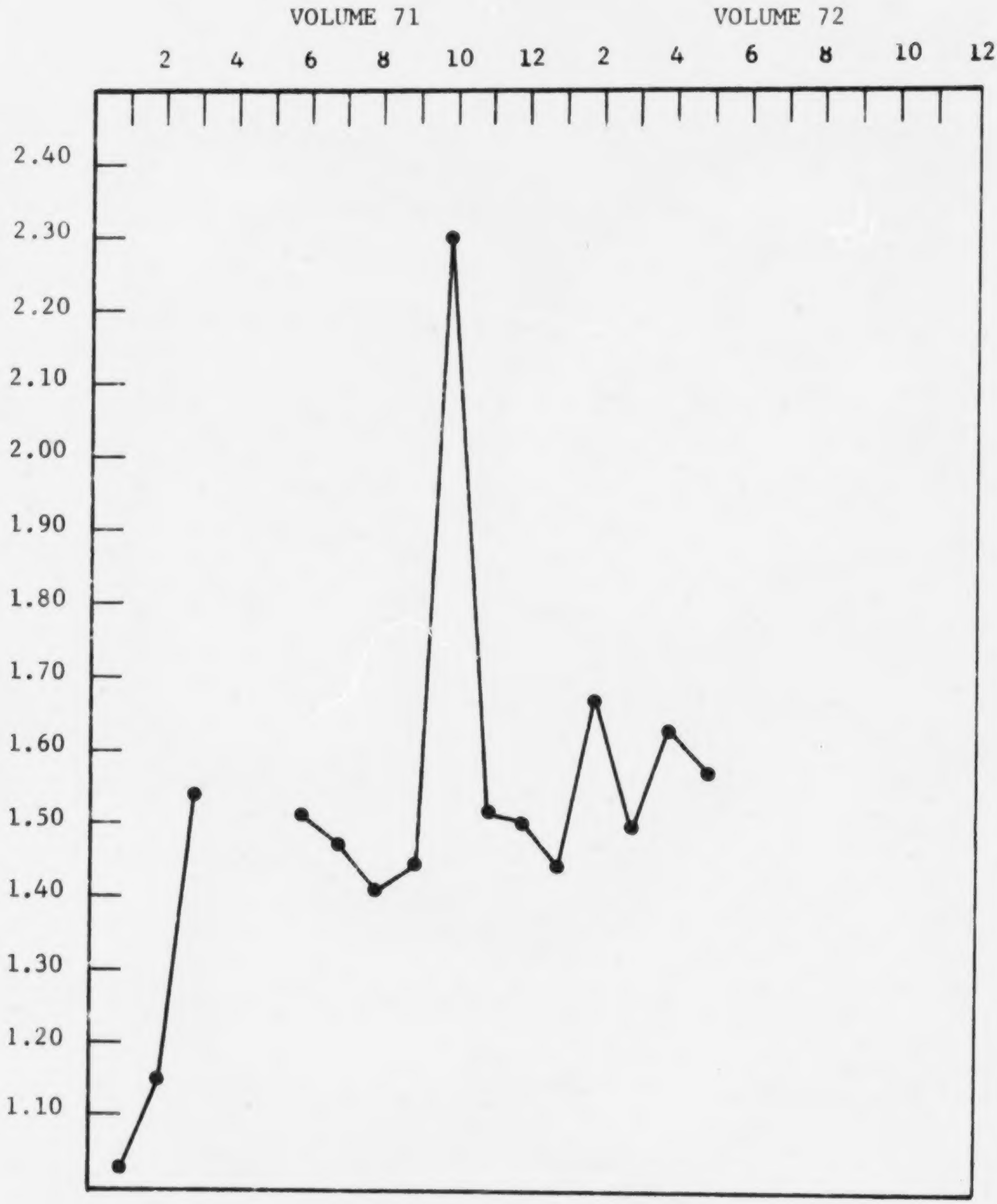


Figure 11-78

HITS PER RETRIEVED CITATION VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 71

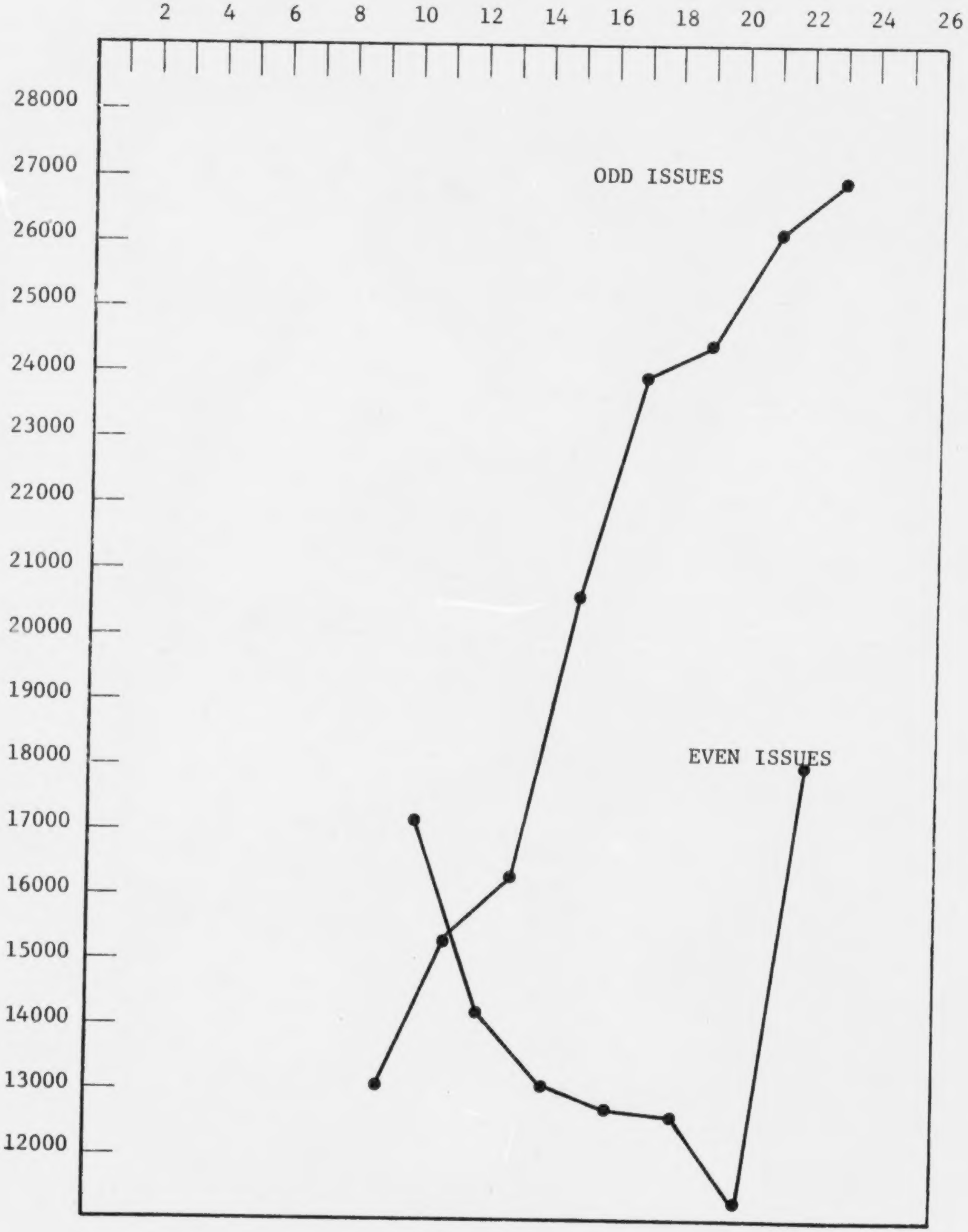


Figure 11-79

NORMALIZED HITS VS. ISSUE

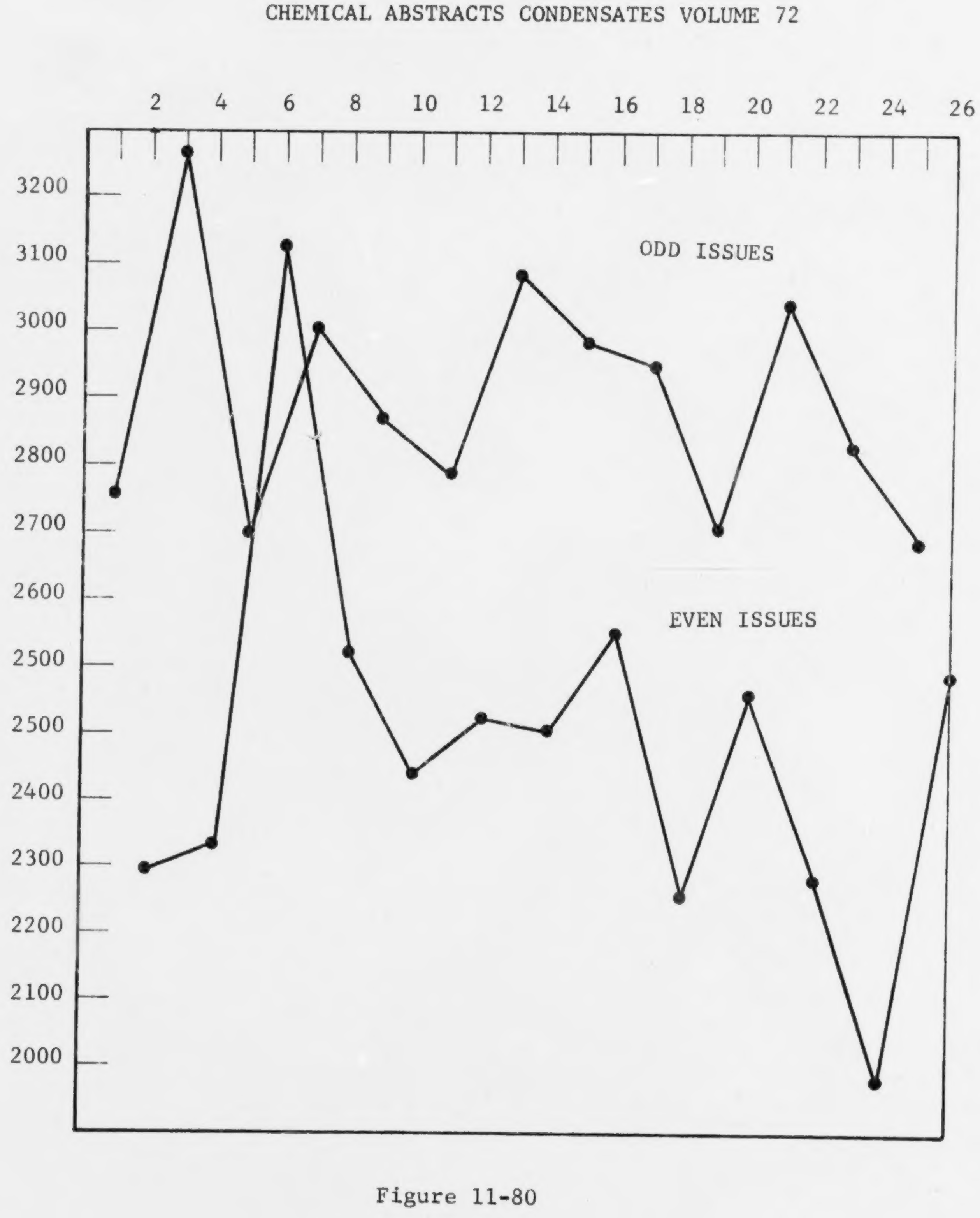


Figure 11-80

NORMALIZED HITS VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

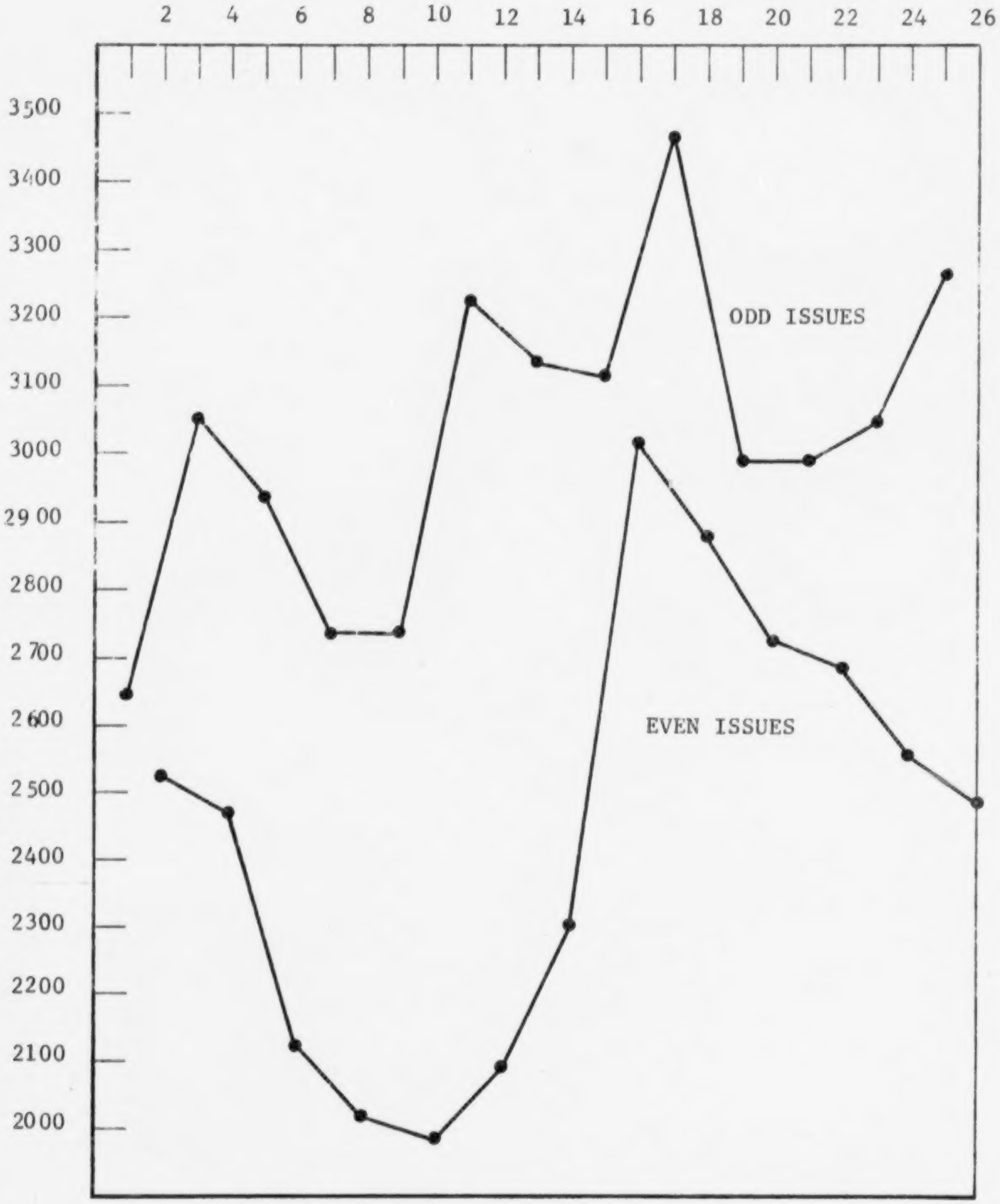


Figure 11-81

NORMALIZED HITS VS. ISSUE

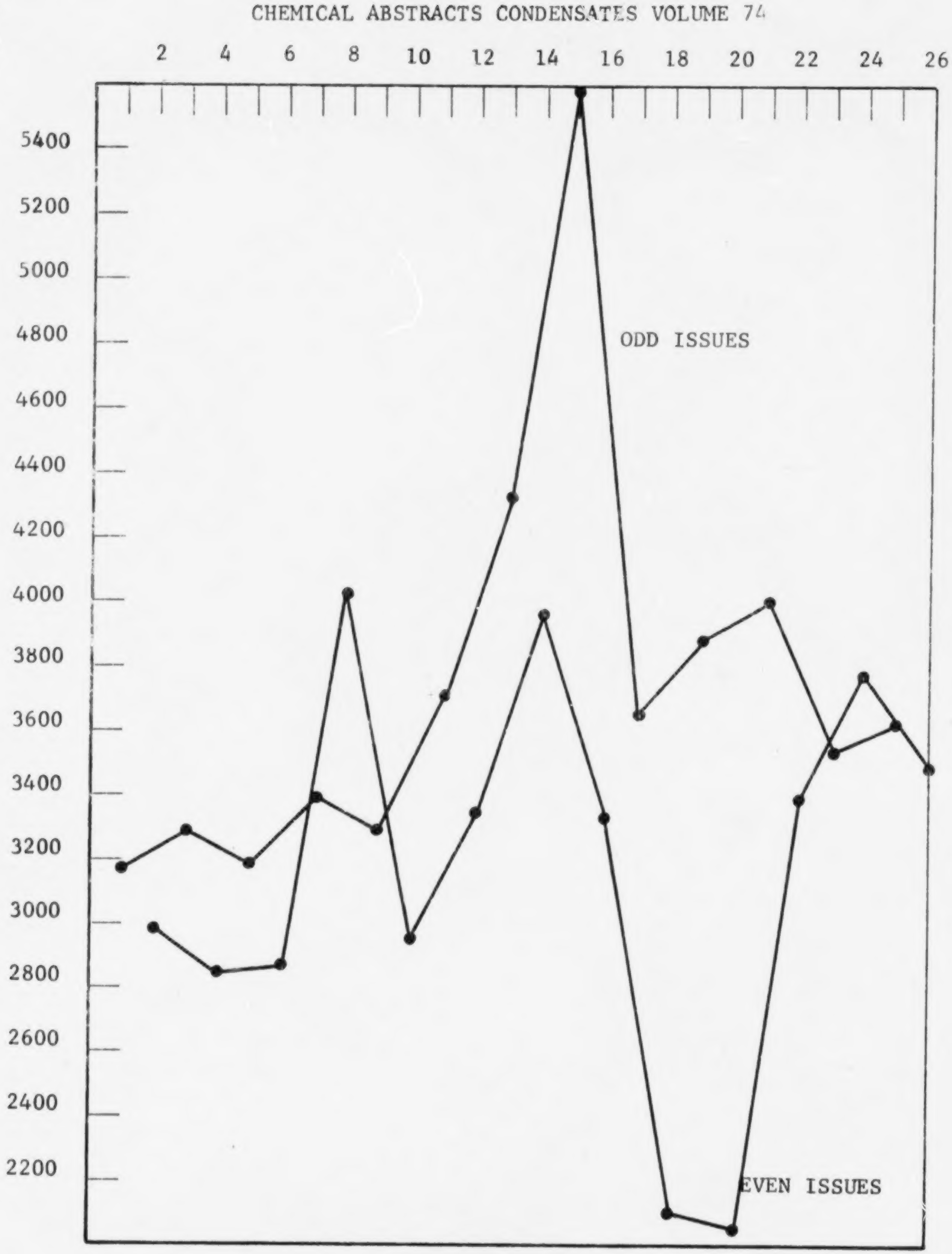


Figure 11-82

NORMALIZED HITS VS. ISSUE

404

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

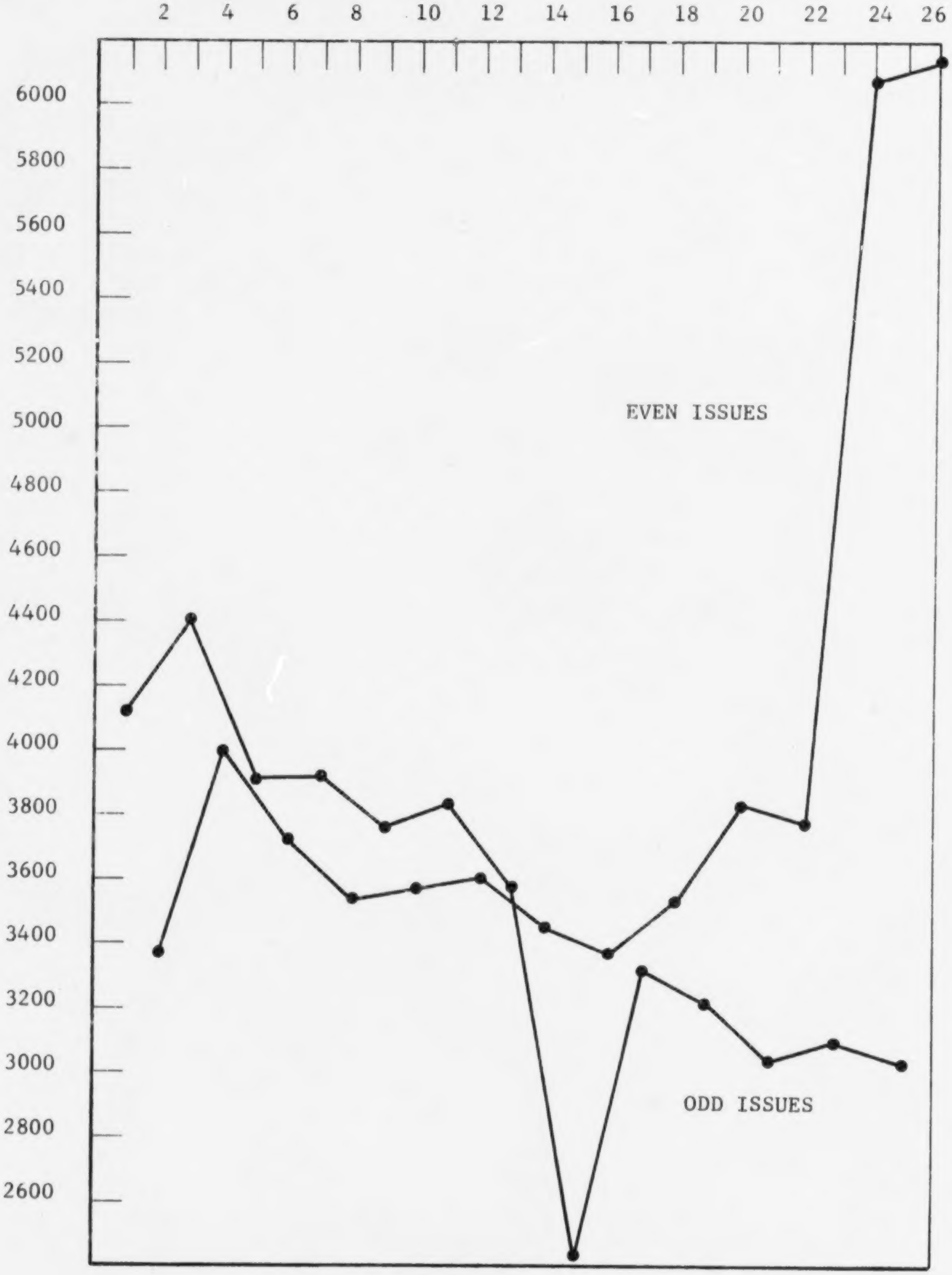


Figure 11-83

NORMALIZED HITS VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

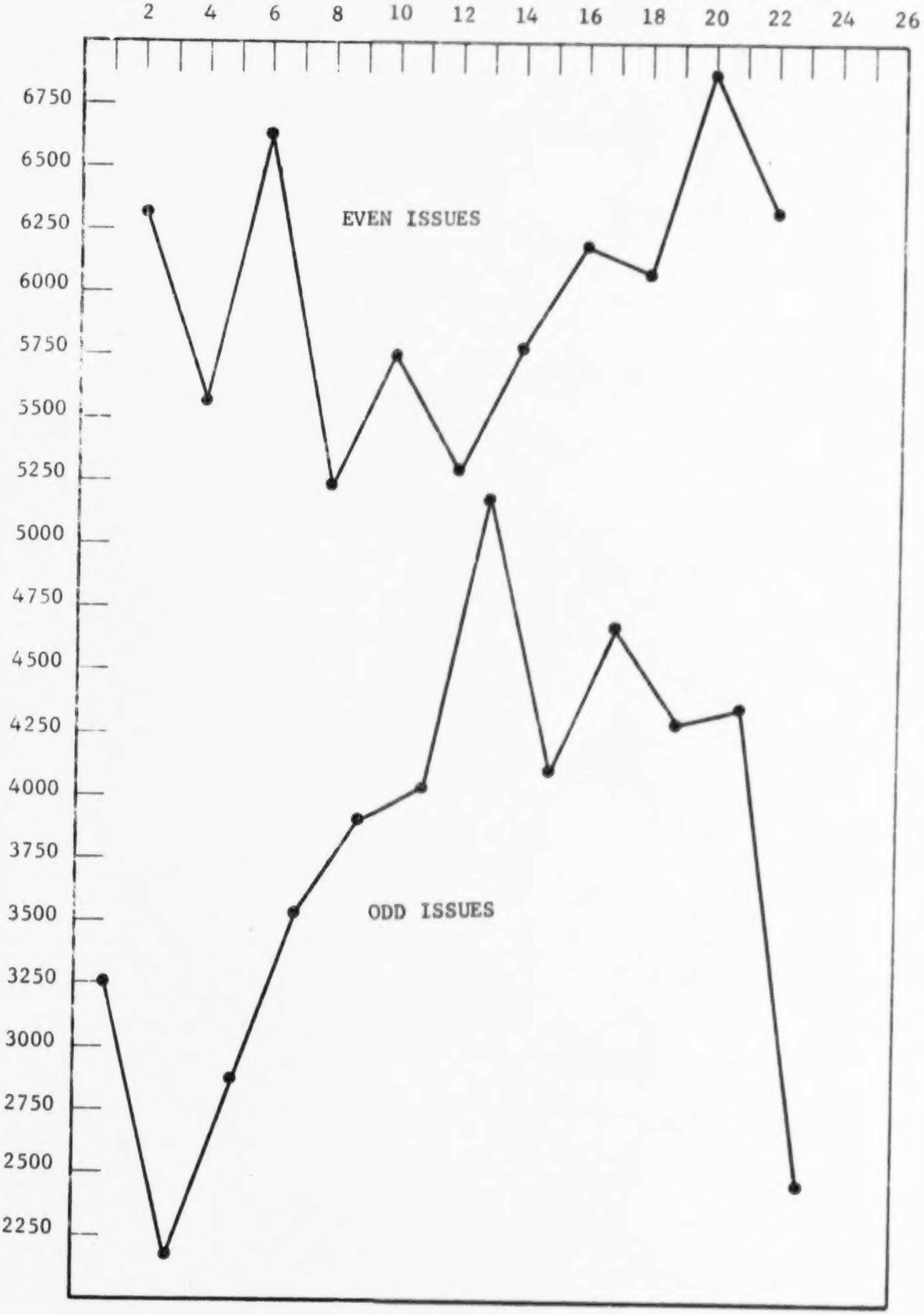


Figure 11-84

NORMALIZED HITS VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUMES 71, 72

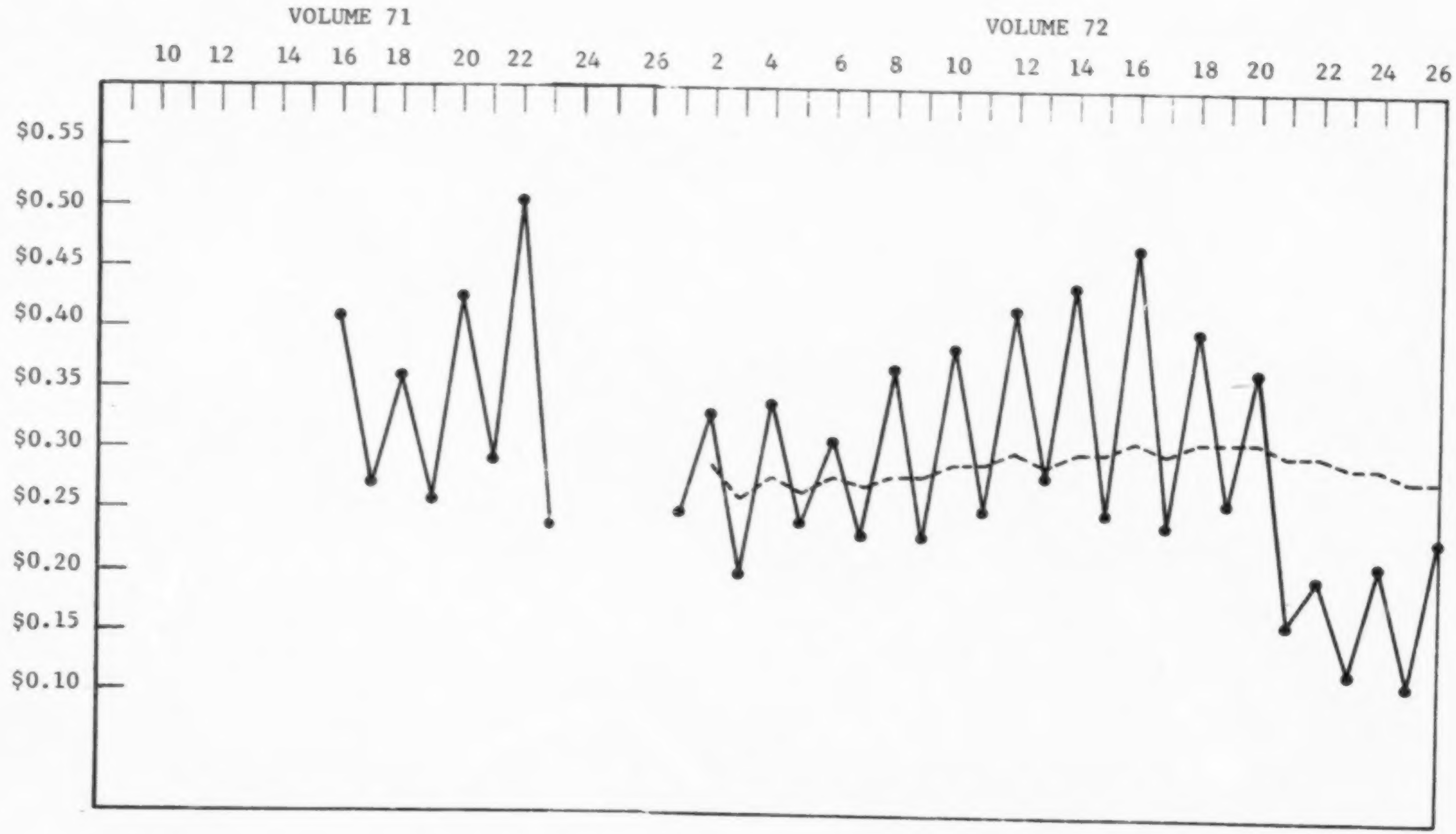


Figure 11-85

COST PER HIT VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 73

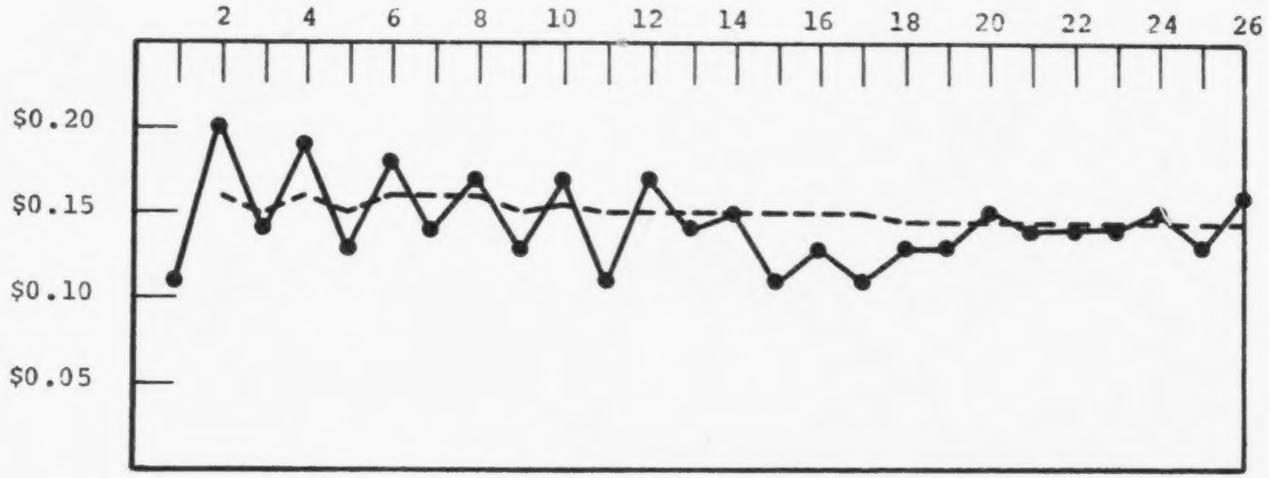


Figure 11-86

COST PER HIT VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 74

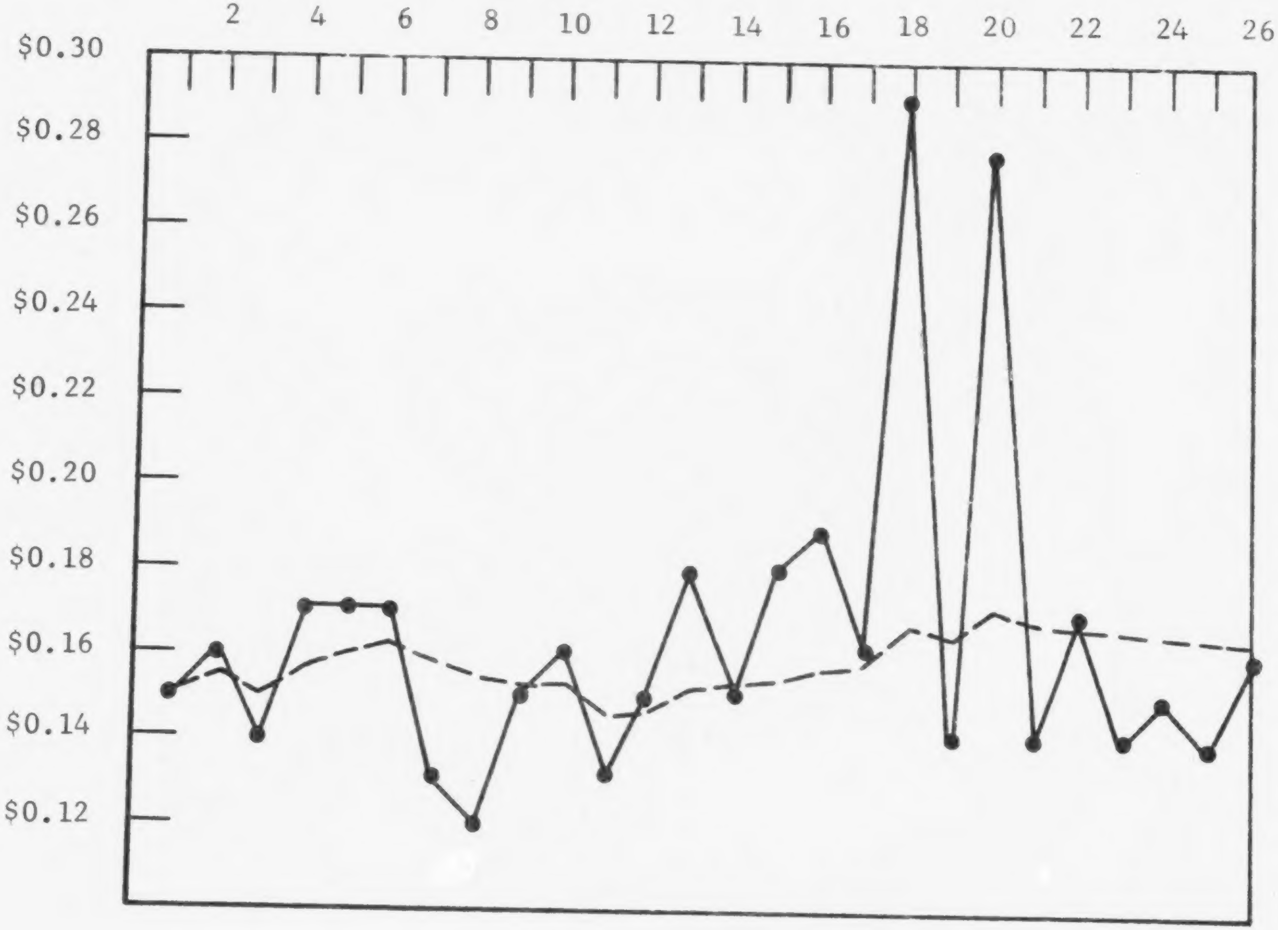


Figure 11-87

COST PER HIT VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 75

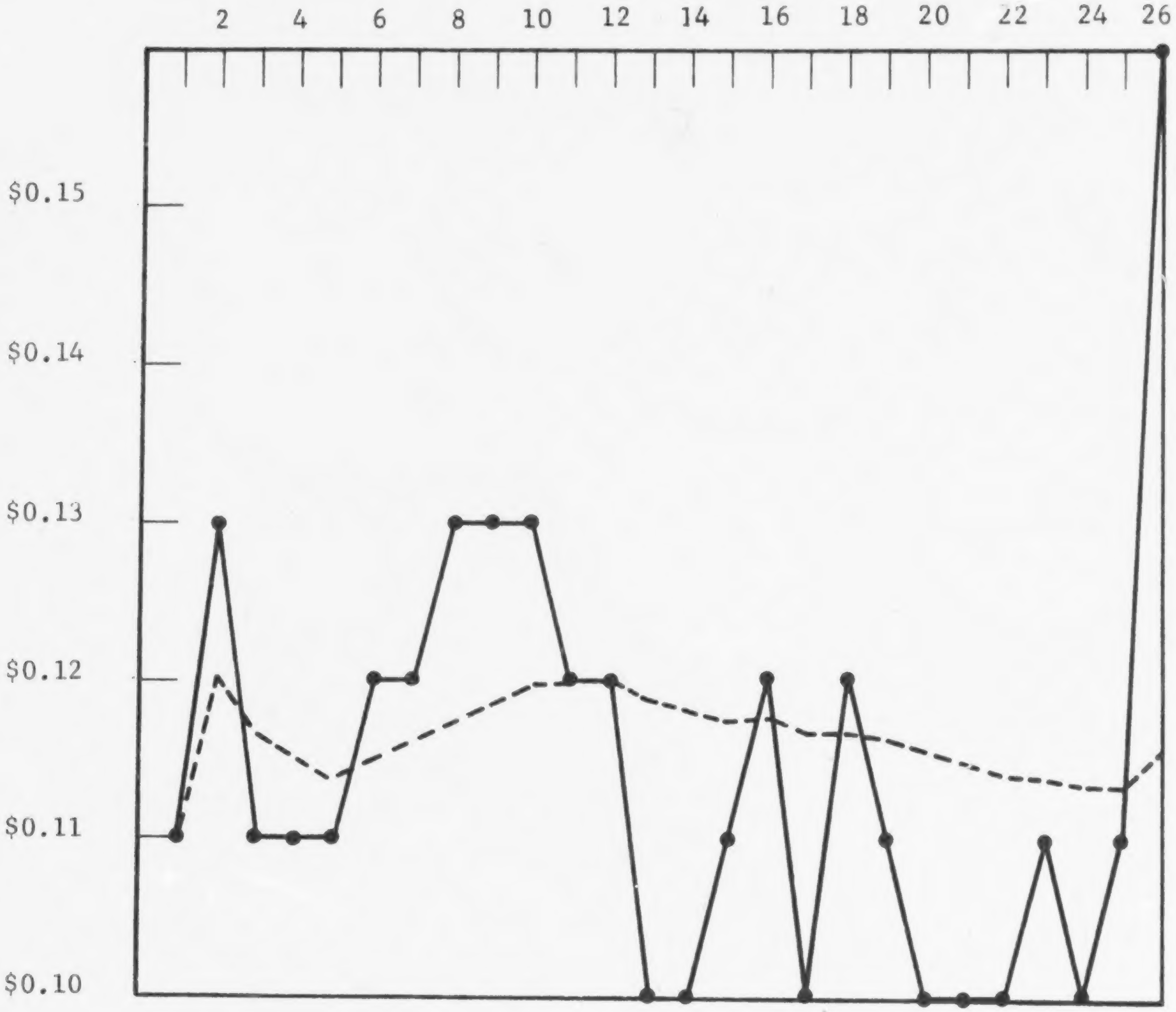


Figure 11-88

COST PER HIT VS. ISSUE

CHEMICAL ABSTRACTS CONDENSATES VOLUME 76

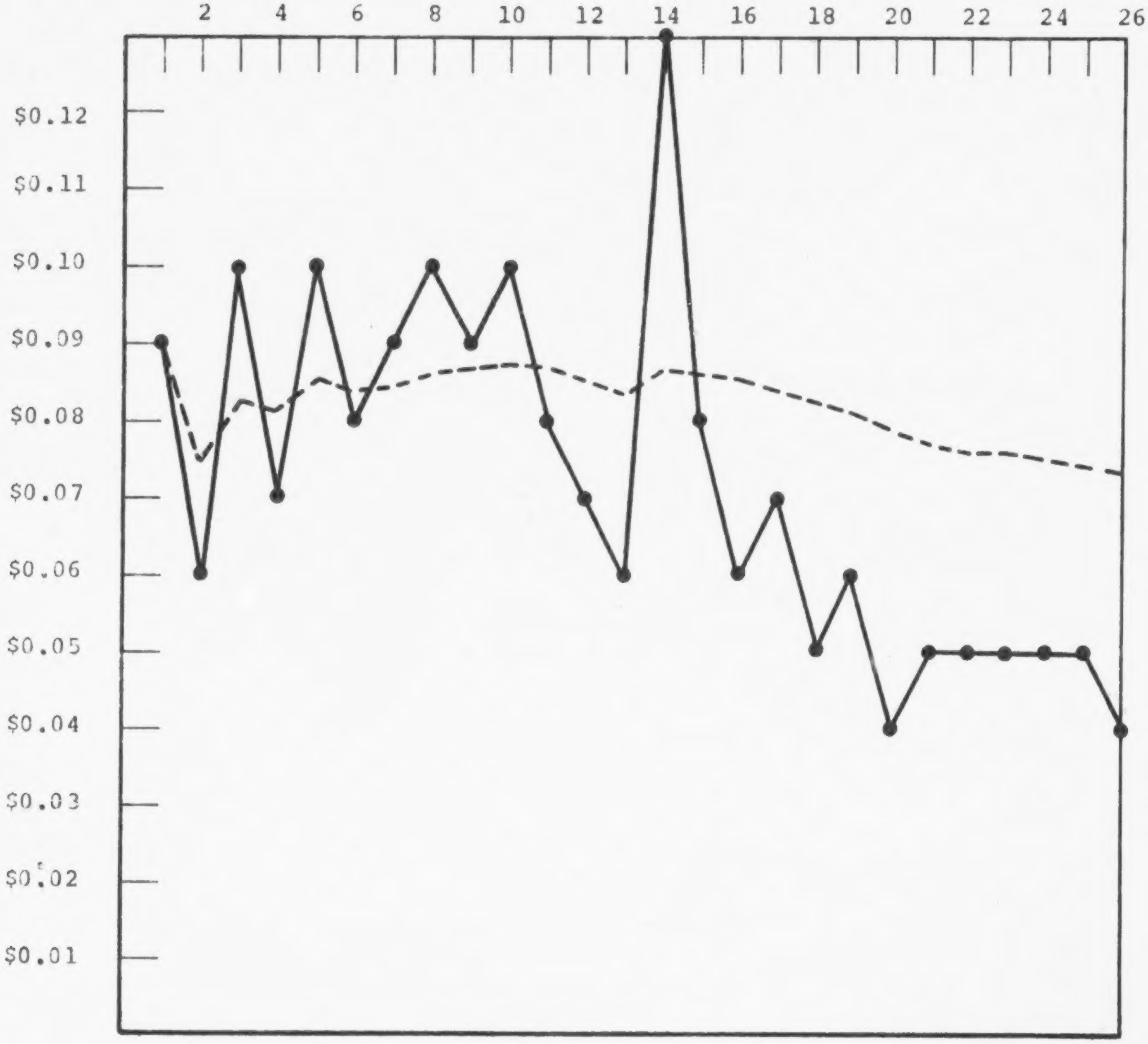


Figure 11-89

COST PER HIT VS. ISSUE

BIORESEARCH INDEX VOLUMES 70, 71

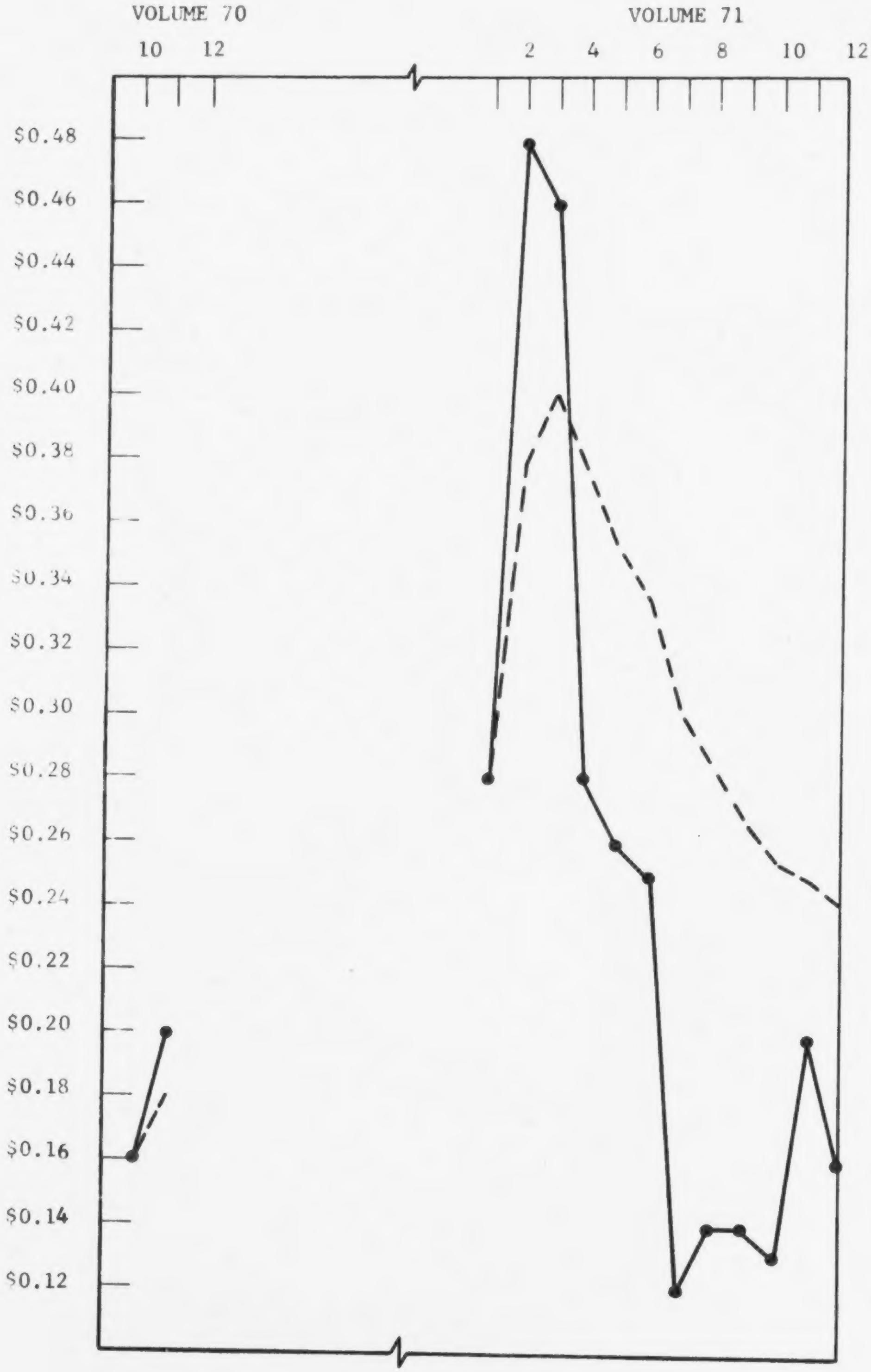


Figure 11-90

COST PER HIT VS. ISSUE

BIOLOGICAL ABSTRACTS PREVIEWS VOLUME 51

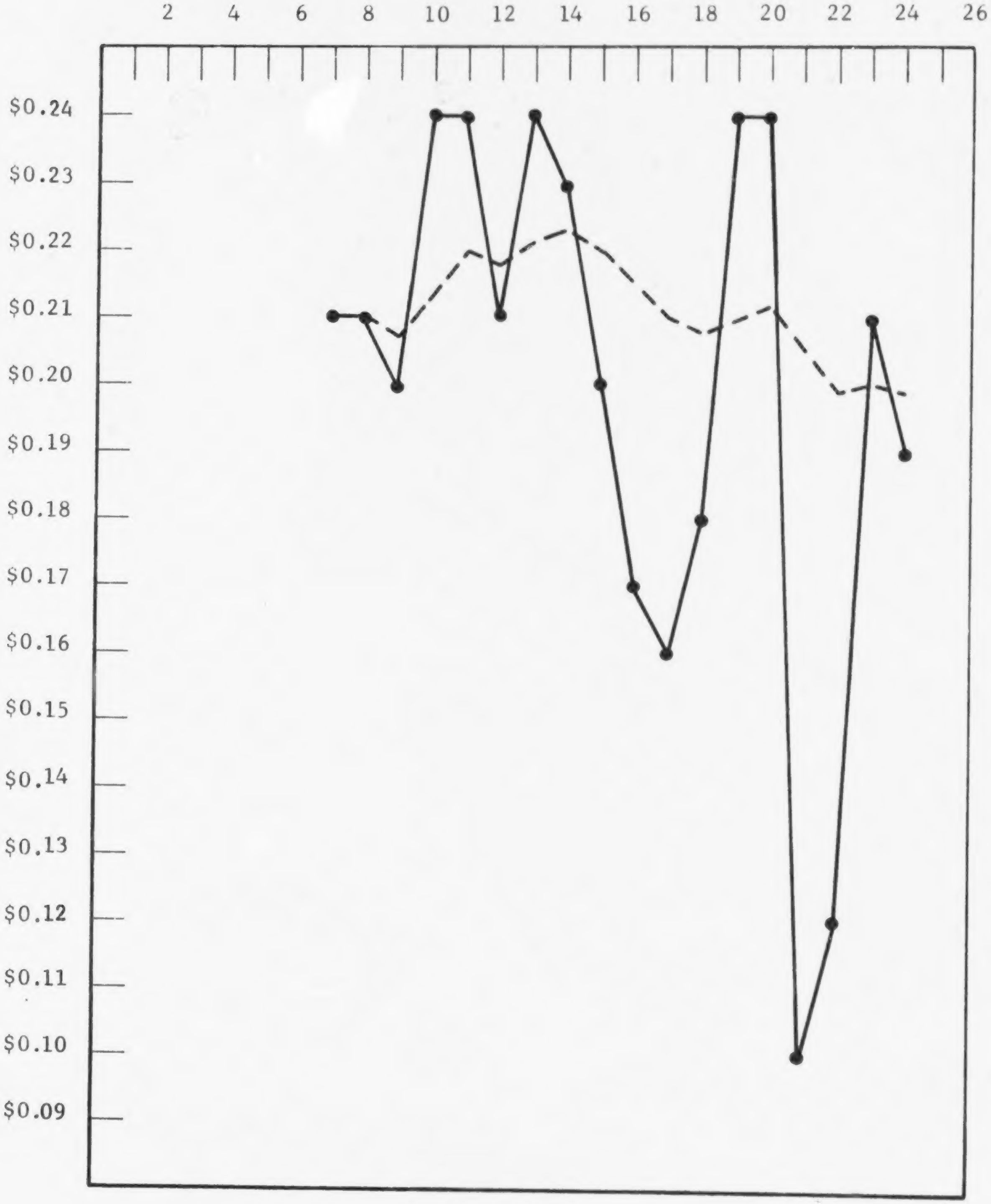


Figure 11-91

COST PER HIT VS. ISSUE

BIOLOGICAL ABSTRACT PREVIEWS VOLUME 52

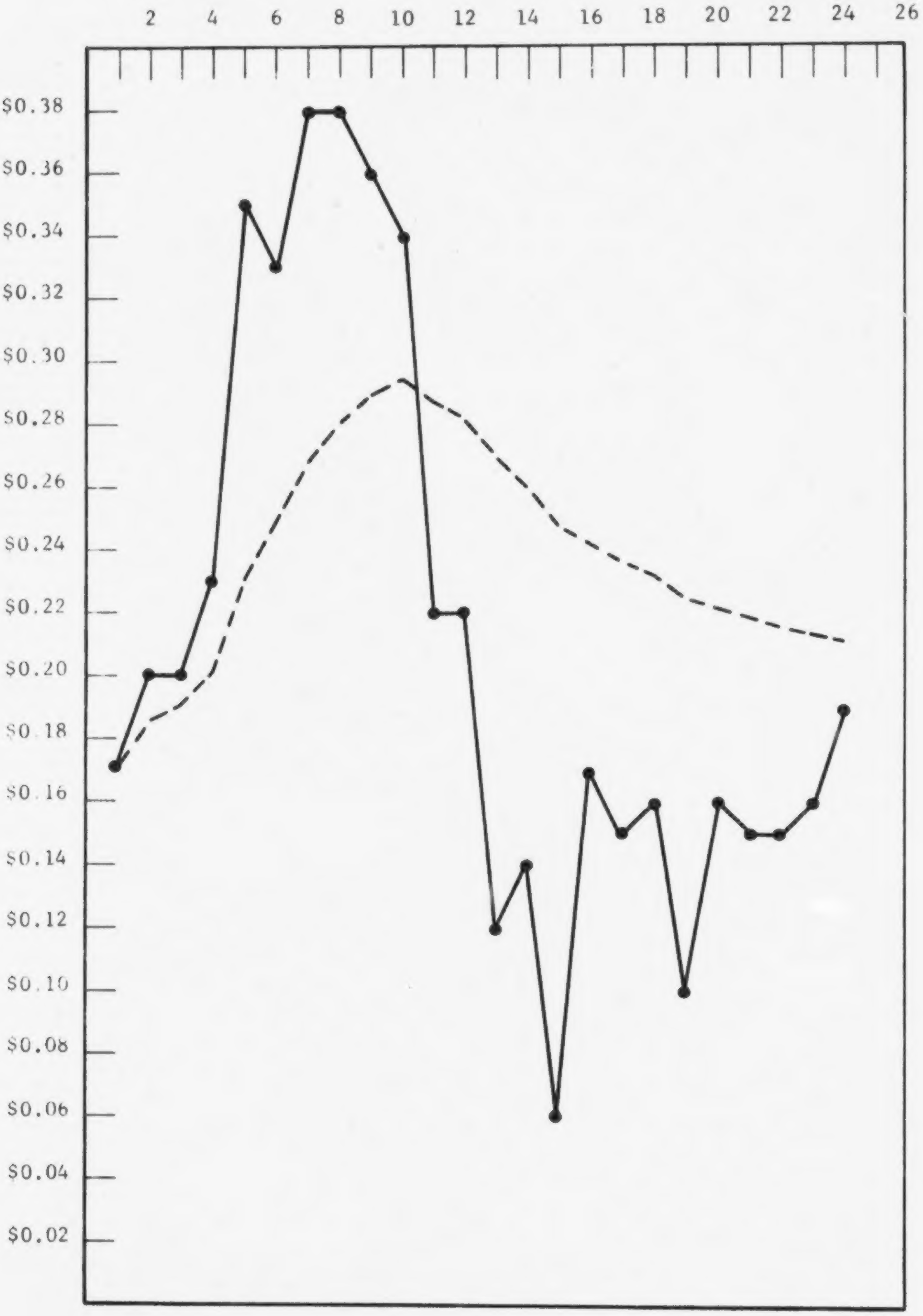


Figure 11-92

COST PER HIT VS. ISSUE

ENGINEERING INDEX COMPENDEX VOLUMES 71, 72

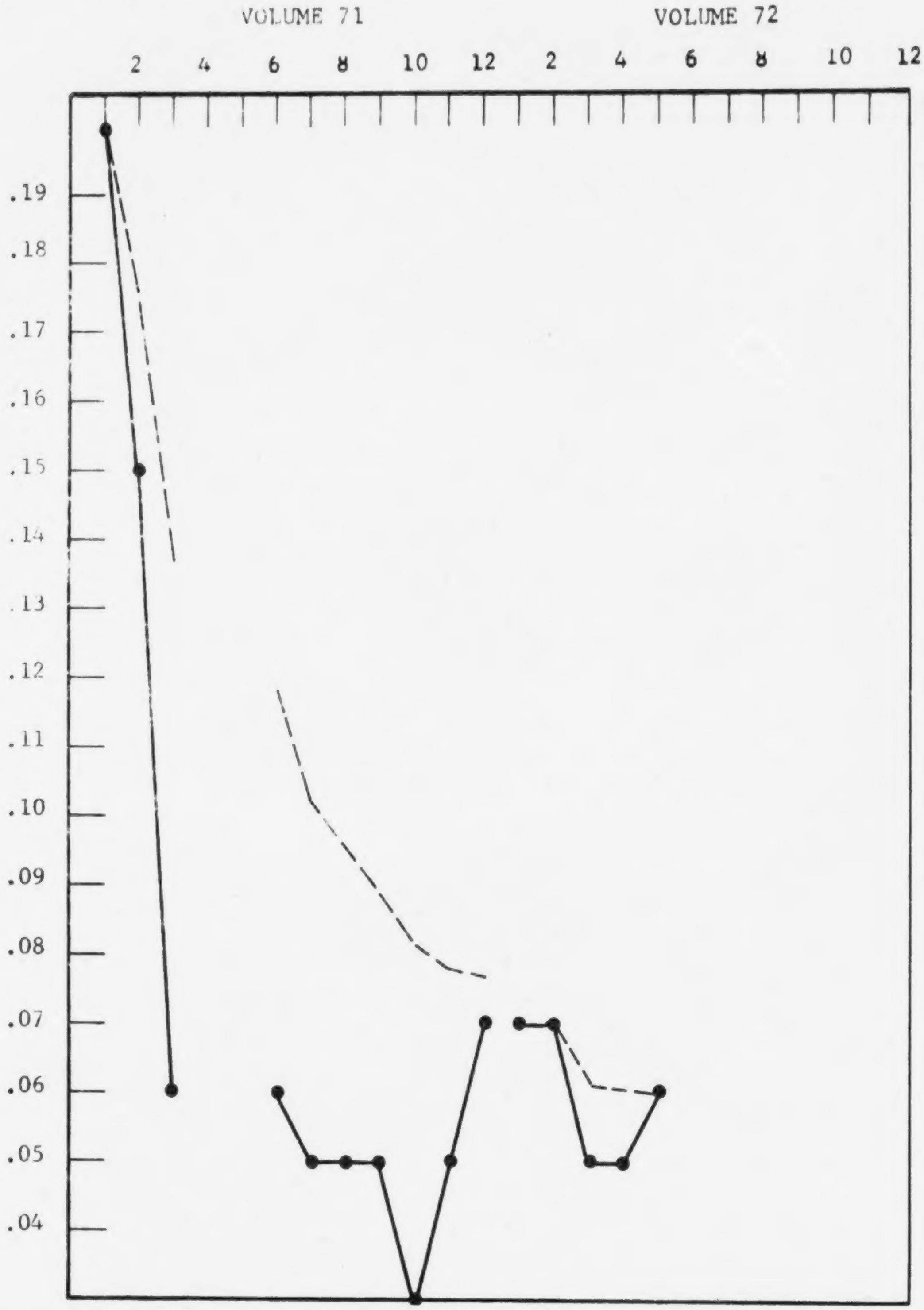


Figure 11-93

COST PER HIT VS. ISSUE

12. CONFERENCES, PRESENTATIONS, PUBLICATIONS, AND PROFESSIONAL ACTIVITIES

Computer Search Center personnel have participated extensively in the professional concerns of the information community. Activities in this field have proven to be a valuable source of two-way communication between the Computer Search Center and other information-processing organizations. For example, a follow-up study of the November 1969, joint meeting of the Chicago Sections of the ACS--Division of Chemical Literature, SLA, and ASIS has shown that of the 54 organizations that attended, 21 have become either trial users or subscribers of CSC. Additional beneficial contacts have resulted from the activities described below. The items below are arranged by professional organization, and within organizations are listed offices held, followed by presentations and publications in chronological order.

American Chemical Society

Publications and talks:

Fanta, P.E., Schwartz, E.S., and Williams, M.E., "Modern Techniques in Chemical Information," presented at the Third Great Lakes Regional Meeting of the American Chemical Society at Northern Illinois University, DeKalb, Illinois, June 5, 1969.

Williams, M.E. and Schipma, P.B., "Design and Operation of a Computer Search Center for Chemical Information," presented at the American Chemical Society meeting in September 1969 and published in the Journal of Chemical Documentation, Vol. 10, No. 3, September 1970.

Williams, M.E., "Information Sciences at IIT Research Institute," seminar for a Joint Meeting of the Chicago Chapter of the American Chemical Society, Special Libraries Association, and the American Society for Information Science, November 1969.

"Searching the Scientific Literature by Computer," exhibit at the September 1970 ACS meeting.

Williams, M.E., "Computer Based Information Retrieval for Chemists," presented at the Rock River Valley Chapter of the American Chemical Society, Rockford, Illinois, March 25, 1971.

Williams, M.E., "Linguistic Aids for Searching Data Bases," presented at the 3rd Central Regional Meeting of the ACS, Cincinnati, Ohio, June 8, 1971.

Williams, M.E., "Computer Search Center," presented at the 5th Great Lakes Regional Meeting of the ACS, Bradley University, Peoria, Illinois, June 11, 1971.

American Society for Information Science

Offices Held:

Williams, Martha E.	Councilor-at-Large, 1971-72 Publications Committee, 1971-72 Chairman, Committee on Inter-Society Cooperation, 1972

Publications and talks:

Schipma, P.B., Williams, M.E., and Shafton, A.L., "Comparison of Document Data Bases," Journal of the American Society for Information Science, Vol. 22, No. 5, September-October 1971.

Preece, S.E., "Data Base Support for an SDI System," presented at 1st Annual Mid-Year Regional Conference of the American Society for Information Science, Dayton, Ohio, May 18-20, 1972.

Schipma, P.B., "PL/1 as an Information Retrieval Language," presented at the First Annual Mid-Year Regional Conference of the American Society for Information Science, Dayton, Ohio, May 18-20, 1972.

Stewart, A.K. and Williams, M.E., "International Information Transfer and SDI," submitted as a contributed paper at the 1972 ASIS annual meeting, Washington, D.C.

Association for Computing Machinery

Offices held:

Williams, Martha E.	Publications Board, 1972-73

Publications and talks:

Onderisin, E.M., "The Least Common Bigram: A Dictionary Arrangement Technique for Computerized Natural-Language Text Searching," presented at the 1971 ACM National Conference, August 3-5, 1971, and published in the Proceedings.

Association of Scientific Information Dissemination Centers

Offices held:

Schipma, Peter B.	Standards Committee, 1969-1972
Schwartz, Eugene S.	President, 1969-1970
Williams, Martha E.	Vice-President, 1971-1972, 1972-1973 Chariman, Cooperative Data Management Committee, 1970-1972 Committee on Center Supplier Relations, 1971-1972

Publications and talks:

Williams, M.E., "The Information Center of 1975," presented at the Association for Scientific Information Dissemination Centers meeting in Atlanta, Georgia, March 1970.

Schipma, P.B., "Term Fragment Analysis for File Inversion," presented at the NFSAIS/ASIDIC Joint Meeting, Washington, D.C., February 23, 1971.

Schwartz, E.S., "The Information Process: Relationships, Problems and Limits," presented at the NFSAIS/ASIDIC Joint Meeting, Washington, D.C., February 23, 1971.

Williams, M.E., "Cooperative Data Management for Information Centers," presented at the NFSAIS/ASIDIC Joint Meeting, Washington, D.C., February 24, 1971.

Williams, Martha E. and Stewart, Alan K., "ASIDIC Survey of Information Center Services." June 1972.

National Academy of Sciences, National Research Council, Committee on Chemical Information

Offices held:

Williams, Martha E.	Committee Member, 1970-1972 Chairman, Large Data Base Subcommittee, 1971-1972

Publications and talks:

Presentation and discussion concerning the Computer Search Center at the January 13-14, 1972 meeting held at Chicago, Illinois.

Large Data Base Survey, 1972.

National Federation of Science Abstracting and Indexing Services

Williams, M.E., "Computer Based Services," seminar presented at the National Federation of Science Abstracting and Indexing Services, New York, April 27-29, 1970 and Cleveland, Ohio, May 25-27, 1970.

Schipma, P.B., "Technological Aspects of Computer Based Services," presented at Seminar of the National Federation of Science Indexing and Abstracting Services in Chicago, May 10-11, 1971.

Williams, M.E., "Information Center--Case History," presented at the NFSAIS Computer Based Services Seminar, Chicago, Illinois, May 10-11, 1971.

Williams, M.E., "Case History--IITRI," presented at the NFSAIS Indexing in Perspective Seminar, Chicago, Illinois, May 24-26, 1971.

Schipma, P.B., "Technological Aspects of Computer Based Services," presented at Seminar of the National Federation of Science Indexing and Abstracting Services in New York, February 3-4, 1972.

Miscellaneous

Publications and talks:

"Computer Search Center," Science Information Notes, Vol. 1, No. 3, May-June 1969, pp. 107-110.

Williams, M.E., "An Information Retrieval System," presented at the American Management Association seminar on Fundamentals of Information Retrieval Systems and Techniques, San Francisco, California, June 5-7, 1968.

Williams, M.E., "The Information Problem," presented at the Institute on Information Resources, Networks, and Retrieval, Department of Engineering, University of Wisconsin, Madison, Wisconsin, November 11-12, 1968.

Schwartz, E.S., "Heuristic Retrieval: Variable Search Strategies for Identification," Journal of Chemical Documentation, Vol. 9, No. 1, 1969, pp. 31-46.

Schwartz, E.S. and Williams, M.E., IIIT Research Institute) and Fanta, P.E., (Illinois Institute of Technology), "Modern Techniques in Chemical Information (Workbook and Syllabus)," February 1969. To be published.

Williams, M.E., "Content Analysis of Documents: An Analytic View," presented at the American Management Association seminar on Fundamentals of Information Retrieval Systems, San Francisco, California, June 21-25, 1969.

Williams, M.E., "Computer Search Center--A One Stop Information Center for Chemical Librarians," presented at the Chemists' Club Symposium, New York, April 9, 1970.

Williams, M.E., "Design of Data Base Systems and Identification of Cost Elements," presented at EDUCOM meeting, Boston, Mass., April 15, 1970.

Williams, M.E., "SDI Whither?" presented at the annual Special Libraries Association meeting in Detroit, Michigan, June 9, 1970.

Williams, M.E., "Provision of Information to the Research Staff," Paper No. 46C presented at the American Institute of Chemical Engineers, 63rd Annual Meeting, Chicago, Illinois, December 3, 1970.

Williams, M.E., "New Techniques of Information Handling," Paper No. 14C presented at the American Institute of Chemical Engineers, 63rd Annual Meeting, Chicago, Illinois, December 3, 1970.

Williams, M.E., "Computer Searching of Multiple Machine-Readable Data Bases," presented at the National Library Week Symposium II, Information for the Seventies, Minneapolis, Minnesota, and published in April 20, 1971, MnU Bulletin, Vol. 2, No. 3, July 1971.

Williams, M.E., "Data Base Utilization--Information Center and Related Applications," presented at the Colloquium on Machine-Readable Data Bases--their Creation and Use, sponsored by the School of Library Science, State University of New York, Albany, New York, April 21, 1971.

"Computerized Information Services for Chemists," Chemistry News, Issued by the Chemical Division of IIT Research Institute, May 1971.

Williams, M.E., "Integration of a Processor-Supplied Data Base with a Standard Center-Oriented System," presented at the Chemical Abstracts Services--CA Integrated Subject File User Seminar, Columbus, Ohio, May 24, 1971.

Williams, M.E., "Use of Machine-Readable Data Bases by Scientists and Engineers," presented at the ASEE annual meeting, Annapolis, Maryland, June 24, 1971.

Williams, M.E., "Experiences of IIT Research Institute in Operating a Computerized Retrieval System for Searching a Variety of Data Bases," presented at the 3rd Cranfield International Conference on Mechanized Information Storage and Retrieval Systems, Cranfield Institute of Technology, Cranfield, England, July 20, 1971, published in Information Storage and Retrieval, Vol. 8, No. 2, pp. 57-75, April 1972.

Williams, M.E., "Handling of Varied Data Bases in an Information Center Environment," Proceedings of Conference on Computers in Chemical Education and Research, Northern Illinois University, DeKalb, Illinois, July 23, 1971.

Schipma, P.B., "IITRI's Computer Search Center" presented at Workshop on Indexing and Index Use of the Institute of Paper Chemistry, August 17, 1971.

Williams, M.E., "The IITRI Computerized System for Searching Multiple Data Bases--Analysis of Design Criteria," presented at the INTREX Seminar, MIT, October 28, 1971.

13. REFERENCES

A complete list of papers and presentations made by CSC staff members is given in Section 12. This section contains the papers referenced in earlier sections and a listing of data base documentation.

13.1 Papers Referenced

1. K. D. Carroll (Compiler & Editor): Survey of Scientific Technical Tape Services. AIPID 70-3, ASIS SIG SIG/SDI September 1970.
2. L. Cohan (Editor): Directory of Computerized Information in Science and Technology. Science Associates/International, Inc., New York.
3. M. E. Williams: Cooperative Data Management for Information Centers. Presented at the Association of Scientific Information Dissemination Centers Meeting, Washington, D. C., February 24, 1971.
4. P. B. Schipma, M. E. Williams and A. L. Shafton: Comparison of Document Data Bases. Journal of the American Society for Information Science, Vol. 22, No. 5, September-October 1971.
5. M. E. Williams: Handling of Varied Data Bases in an Information Center Environment. Presented at the Conference on Computers in Chemical Education and Research, Northern Illinois University, DeKalb, Illinois, July 23, 1971.
6. P. B. Schipma: Term Fragment Analysis for Inversion of Large Files. Presented at the Association of Scientific Information Dissemination Centers Meeting, Washington, D. C., February 24, 1971.

13.2 Data Base Documentation

American Institute of Physics
New York, New York

SPIN/O. A Magnetic Tape Service
of the American Institute of Physics

Bio-Sciences Information Service
Philadelphia, Pennsylvania

Guide to the Contents of BA Previews

Chemical Abstracts Service
Columbus, Ohio

Data Content Specifications
for CA Condensates in S.D.F.

Data Content Specifications
for the CA Integrated Subject File in S.D.F.

Data Content Specifications
for Chemical Titles in S.D.F.

Data Content Specifications
for Chemistry Industry Notes in S.D.F.

Data Content Specifications
for Patent Concordance in S.D.F.

Data Content Specifications
for Polymer Science & Technology in S.D.F.

Standard Distribution Format (S.D.F.)
Technical Specifications (revised)

Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia

Clearinghouse Announcement Journal
Available on Magnetic Tape

ERIC Processing and Reference Facility
Bethesda, Maryland

ERIC Master Files, Magnetic Tape Formats

MARC II Format of the ERIC Data Base

INSPEC, The Institute of Electrical Engineers
London, England

Magnetic Tape Files Devices from the INSPEC Data Base

Institute for Scientific Information
Philadelphia, Pennsylvania

ISI Magnetic Tapes

International Food Information Service
Frankfort am Main, Germany

IFIS Magnetic Tape Manual

Library of Congress
Washington, D.C.

Subscriber's Guide to the MARC Distribution Service

14. CONCLUSION AND SUMMARY

The IITRI CSC was begun in July 1968. The first year was spent in design and testing in preparation for providing information services from machine-readable data bases to users on a cost-recovery basis. Over the past three years CSC has profided SDI and retrospective search services to a varied and dispersed group of users in industry, academia, and government. We designed the system to handle virtually any document-type data base--and it does--and the data bases we have used are BA, CA, and EI. We have processed approximately 600 profiles for 2500-3000 people, and we have searched more than 2 million citations ranging from 200-800 characters each. From this experience we have gathered statistical data, analyzed the data, and conducted research. Our findings both verify the design parameters and provide bases for monitoring and improving the overall system--including the data bases, software, profiles, users' reactions, and system operators as well as all of the interfaces between them. The work discussed in this report does not relate to hypothetical cases, research prototypes, or pilot studies. The report discusses what we have done and are doing, plus observations regarding the real life situation of providing services on a production basis to users who pay for the service.

At present we have completed four years work under NSF Contract 554, and a no-cost time extension has been granted for continuing the contract through December 1972. Virtually all of the design research and development work has been completed, and the center is well on the way to becoming self-supporting. The major problems affecting marketing are, on the side of potential users, the lack of awareness and understanding of machine-readable data base and their potential; and on the part of centers, the existence of duplicative efforts and coverage. Through the auspices of ASIDIC, we look forward to resource sharing and informal networking as a means of improving the distribution of the available products to an as yet limited but potentially sizeable market. Machine-readable data bases are here to stay, and they fulfill a real need, but efforts regarding repackaging of data and development of new services from the data bases together with education of potential users is needed.